Curso 2009-2010

Apuntes

de

BIOESTADISTICA

Por Eduardo Buesa Ibáñez, Profesor de la asignatura en la Escuela Universitaria de Enfermería Nª Sª del Sagrado Corazón. Castellón

ESCUELA UNIVERSITARIA DE ENFERMERIA “Nª Sª DEL SAGRADO CORAZÓN”.
HOSPITA GENERAL DE CASTELLON

B I O E S T A D I S T I C A

OBJETIVOS: Realizar una introducción elemental en el campo de la Metodología Estadística para que el futuro Diplomado sea capaz de aplicar los procedimientos estadísticos fundamentales y valorar críticamente los informes y publicaciones que hagan uso de tales métodos.

CONTENIDOS: Temas de Estadística Descriptiva, de Estadística Inferencial y de algunas aplicaciones concretas de la Estadística en las Ciencias de la Salud. El alumno aprenderá a recoger datos procedentes de muestras, a ordenarlos y a presentarlos en forma de tablas, gráficos y números índice que los resumen (media, varianza, desviación estándar, etc). Además aprenderá a estimar parámetros y a realizar pruebas de conformidad, relación y contraste de variables.

METODOLOGIA: Exposición de los temas. Realización de más de 200 ejercicios prácticos. Manejo de programas estadísticos libres y gratuitos.

EXAMENES: Ante todo, resolución de uno o varios supuestos prácticos. Alguna pregunta sobre teoría “tipo test” o a contestar en una o dos líneas.

PROGRAMA:

Tema 1	Fundamentos y fines de la Bioestadística.
Tema 2	Operaciones matemáticas más usuales en Bioestadística.
Tema 3	Variables y su medida. Síntesis de datos estadísticos.
Tema 4	Tabulación de datos.
Tema 5	Representaciones gráficas.
Tema 6	Indices estadísticos de variables cuantitativas. Parámetros de tendencia central, dispersión, posición y forma.
Tema 7	Datos bivariados. Tabulación y representación gráfica. Correlación y regresión.
Tema 8	Series de tiempo.
Tema 9	Teoría de la probabilidad
Tema 10	Distribuciones fundamentales de probabilidad (normal, binomial, de Poisson). Otras distribuciones.
Tema 11	Planificación de estudios estadísticos. Clases de estudios.
Tema 12	Recogida de la información. Técnicas de muestro. Errores de los muestreos.
Tema 13	Intervalos de probabilidad y confianza. Hipótesis y decisiones estadísticas.
Tema 14	Estimación de parámetros. Pruebas de conformidad
Tema 15	Pruebas de contraste de variables.
Tema 16	Contraste de dos variables cualitativas. Odds ratios.
Tema 17	Contraste de una variable cualitativa y otra cuantitativa.
Tema 18	Contraste de dos variables cuantitativas.
Tema 19	Demografía sanitaria. Medida de la salud.
Tema 20	Errores de las medidas de laboratorio. Control de calidad. Valoración de pruebas diagnósticas
Tema 21	Programas para resolver problemas estadísticos.
Tema 22	La Estadística en Internet

Libros de consulta recomendados

- ESTADISTICA PARA LA INVESTIGACION BIOMEDICA. P Armitage, G Berry. Edit. Doyma, Barcelona
- BIOMETRÍA. RR Sokal, FJ Rohlf. Ediciones Blume, Madrid.
- ESTADISTICA. Gilbert. Ed. Interamericana, Madrid
- ESTADISTICA PARA BIOLOGIA Y CIENCIAS DE LA SALUD. JS Milton. Edit. McGraw-Hill, Madrid

Tema 1 : Fundamentos y fines de la Bioestadística

--Conceptos básicos

La **BIOESTADISTICA** es la Estadística aplicada a las ciencias biológicas.
La **ESTADISTICA** es muy difícil de definir. Esto hace que haya muchas definiciones y que incluso algunos libros la soslayen. Una definición aceptable es **:"La Estadística es el estudio científico de datos numéricos referidos a características variables".**
Un estudio es científico si utiliza métodos rigurosos en su concepción y desarrollo, teniendo como normas básicas la objetividad, el espíritu crítico y la ética. Algunas afirmaciones aparentemente científicas no lo son al no cumplir alguna de estas normas básicas. Es frecuente cuando se tocan temas religiosos, políticos o económicos. Incluso los muy expertos en una materia no están libres de prejuicios y presiones crematísticas.
Los datos numéricos son números que expresan medidas (datos métricos) o recuentos de modalidades (datos categóricos).
Por característica se entiende una propiedad o condición claramente reconocible en diversos individuos. El individuo es la unidad estadística y puede ser una persona, un animal, una planta, un objeto o una acción. Las características pueden ser constantes o variables.
Las *constantes* no varían, siempre ocurren de la misma forma, como las constantes físicas o la certeza de la muerte en los seres vivos. Siguen el llamado modelo determinista de los fenómenos naturales. Tienen un resultado fijo, que se puede resumir por una fórmula matemática. Al lanzar una bola es posible saber con exactitud la velocidad y la aceleración que va a tener en un determinado momento.
Las *variables* presentan una gama de variaciones (al menos dos) en los diversos individuos, como el sexo o la talla de las personas. Siguen el modelo indeterminista (= probabilístico, casual o estocástico). No tienen un resultado fijo. Hay un conjunto de posibles resultados, conocidos de antemano, de los que sólo se producirá uno. Los factores que influyen en que se produzca ese resultado u otro son múltiples, complejos, incontrolables y en parte desconocidos, de forma que el resultado ocurre de forma aparentemente casual, al azar. El azar no es ciego, tiene sus modelos de comportamiento, predecibles con un margen de variación mediante fórmulas matemáticas, basadas en el *cálculo de probabilidades*. Son las llamadas distribuciones fundamentales de probabilidad (Distribución normal, de Poisson, binomial, hipergeométrica, etc.). Los fenómenos biológicos siguen uno u otro modelo, que una vez conocido nos permite calcular las probabilidades de que ocurra tal o cual resultado. ¡EL AZAR ES LA SUPREMA LEY DE LOS FENÓMENOS BIOLÓGICOS!.
En Estadística sólo interesan las características variables, que habitualmente son denominadas variables, sin más aditamentos.

--Etimología e Historia

Estadística proviene de Estado, ya que fueron los Estados los que iniciaron la recogida de datos para su mejor funcionamiento (impuestos, soldados...). Así, hay constancia histórica de censos de tierras y hombres en Egipto 3000 años A.C., en China 2200 años A.C. y en Israel (Moisés y David, 1500 y 1000 años A.C.). En los Evangelios se dice que Jesús nació cuando su familia se trasladaba para cumplimentar el censo ordenado por el César. Por este origen se han introducido términos "humanos" en el lenguaje estadístico, como individuo y población.
Esta Estadística era muy elemental, fundamentalmente recuentos. A partir del siglo XVII experimenta un gran impulso, que se intensifica en siglos posteriores. Se hace científica. En este desarrollo hay que destacar como motores importantes:
1. Los juegos de azar, sobre todo el de dados, que fascinaron a matemáticos insignes y de cuyo estudio nació la teoría de la probabilidad.
2. La Astronomía, con su interpretación de observaciones, cuantificación de posibles errores de medida y predicción de eventos.

3. La Agricultura, con sus estudios genéticos y de productividad.
4. Las compañías de Seguros norteamericanas, con sus estadísticas vitales y estudios de supervivencia y de los factores que más influyen en la misma (edad, tensión arterial, obesidad...)
Nombres como De Moivre, Bernouilli, Lagrange, Laplace, Gauss, Pascal, Quetelet, Galton, Spearman, Pearson y Fisher ocupan un lugar destacado en el progreso de la Estadística.

POBLACIONES Y MUESTRAS

Población: todos los individuos que poseen una determinada característica.
Por su tamaño las poblaciones pueden ser finitas o infinitas. En la práctica, y para facilitar los cálculos, una población se considera "infinita" a partir de un tamaño de 10.000 individuos.
La obtención de datos de una población se llama censo.
Teóricamente un individuo puede tener infinitas características y por tanto puede formar parte de infinitas poblaciones.
Muestra: es una parte de la población, un subconjunto de la misma. Cuando la muestra es representativa de la población, se pueden hacer extensivos a la población los resultados obtenidos en la muestra. En el tema 12 se estudian las muestras con detalle. Aquí se puede adelantar que la representatividad, el que la muestra reproduzca lo más fielmente posible a la población de la que procede, depende fundamentalmente de dos factores: un tamaño adecuado y la elección de los individuos al azar.
Un conjunto de individuos, según las circunstancias, puede ser población o muestra. Por ejemplo, los alumnos de esta Escuela serán "población" cuando tomemos a unos cuantos de ellos para estimar la talla de todo el alumnado de la Escuela. Y serán "muestra" si toda la Escuela ha sido seleccionada para participar en un estudio a nivel nacional.

Hay muchos sinónimos para los conceptos estadísticos:
Bioestadística: Biometría, Estadística biológica...
Población: universo, colectivo, conjunto...
Individuo: elemento, sujeto, efectivo, caso...
Dato: observación, registro, resultado...

CLASES DE ESTADISTICA

Hay que distinguir entre Estadística descriptiva y Estadística inferencial.
La E. descriptiva es la parte más antigua y la más conocida por los profanos. Comprende la obtención, clasificación y presentación de datos numéricos mediante tablas, gráficos, frecuencias, porcentajes, etc. . La vida diaria está invadida por estadísticas de este tipo: de consumo, producción, accidentes, desempleo, etc.
La E. inferencial (o deductiva) es la parte más moderna y científica. A partir de una muestra representativa permite sacar conclusiones razonablemente válidas para la población de origen (Problemas de estimación). Además permite contrastar variables (Problemas de contraste) y concluir si las diferencias o relaciones observadas son explicables o no por el azar.
La E. inferencial *clásica* proporciona un conjunto de "recetas" para realizar las inferencias. Modernamente se ha desarrollado con bastante éxito una variante, la E. *bayesiana*, que se basa en probabilidades condicionadas y que es la base del diagnóstico por computadora.

LA ESTADISTICA, ¿CIENCIA INEXACTA?

Aunque utiliza herramientas matemáticas, las conclusiones estadísticas no son dogmáticas. Incluyen un margen de variación (el llamado intervalo de confianza) y un grado de fiabilidad (nivel de aceptación o significación). Si se estudia por medio de una muestra la opinión de la población de Castellón sobre un determinado asunto y se encuentra que al 65% le parece bien, la Estadística dirá que el 65% está a favor , pero añadirá que este resultado tiene un margen de variación

del, digamos, 10% por encima y debajo de ese valor puntual obtenido y que esta afirmación se hace con una probabilidad de acierto del 95% (o probabilidad de error del 5%).
Es importante destacar que las conclusiones de los estudios estadísticos inferenciales son válidas a nivel de grupo. A nivel individual pueden no serlo por la existencia del llamado error muestral, que suele ser muy pequeño, pero nunca cero. Ejemplo: el medicamento A es eficaz en el 95% de los pacientes con la enfermedad X; el medicamento B sólo en el 5%. Un estudio estadístico permitirá sin duda concluir que el medicamento A es el de elección. La inmensa mayoría se curará sólo con el A. Pero habrá pacientes, pocos ciertamente, que se curen con el B y no con el A.
En la vida diaria se abusa mucho de expresiones como "estadísticamente demostrado" o "estadísticamente comprobado". En realidad la Estadística no demuestra nada, sino que apoya con la fuerza de una probabilidad una determinada conclusión. Admite siempre una probabilidad de equivocarse, que aunque sea muy pequeña, ocurrirá de vez en cuando. Es una ayuda para la toma de decisiones razonables en caso de incertidumbre, aportando las probabilidades de éxito y fracaso de una decisión.
Por otra parte la existencia de una correlación entre dos cosas sólo permite establecer una relación de causalidad si se cumplen determinadas condiciones, ya que puede tratarse de correlaciones espurias, a veces difíciles de descubrir. Dos ejemplos: 1) si en una ciudad se comprueba que la venta de música clásica aumenta a la par que los espectadores que acuden al campo de fútbol, sería muy aventurado concluir que la visita de los campos estimula la afición musical clásica 2) Bernard Show destacó que los londinenses que usaban paraguas estaban mejor nutridos, gozaban de mejor salud y vivían más que los que no lo usaban. Sería peregrino pensar que eso se debía al paraguas. Más bien parecía deberse a que en aquellos tiempos los que usaban paraguas eran los ricos, que disfrutaban de una vida más saludable. En los medios de comunicación, en las argumentaciones de los políticos y grupos de presión e incluso en las publicaciones científicas se utilizan de forma mucho más sutil que en los ejemplos anteriores, de forma más o menos consciente, "conclusiones" estadísticas para hacer comulgar al lector u oyente con grandes ruedas de molino. La Estadística es siempre honesta. los que la utilizan a veces no.

DOS OPINIONES ILUSTRES SOBRE LA ESTADISTICA

1. Hay tres clase de mentiras: mentiras, mentiras viles y estadísticas (Disraeli)
2. El buen cristiano debe guardarse de los matemáticos y de los que practican la predicción... porque existe el peligro de que esta gente esté aliada con el diablo. (San Agustín)

...Y OTRA OPINION ALGO MENOS ILUSTRE...

Y todo esto...¿para qué sirve? (Un antiguo alumno de esta Escuela)



Fisher

Tema 2 : OPERACIONES MAS USUALES EN ESTADISTICA

---OPERACIONES

1) Las “4 reglas” clásicas : sumar, restar, multiplicar y dividir.
2) Potenciación : a^n , generalmente a^2. Recordar que $a^0=1$ y $a^1= a$
3) Radicación : casi exclusivamente la raíz cuadrada
4) Resolución de ecuaciones : nosotros sólo veremos de primer grado
5) utilización del sistema de coordenadas rectangulares (x , y), a veces los 4 cuadrantes, pero habitualmente sólo el primer cuadrante.
6) logaritmos y antilogaritmos. Fáciles de obtener con una calculadora científica (log , ln , 10^x , e^x)
7) Factoriales : n! , que es igual a n*(n-1)*(n-2)*(n-3)......*1 . Recordar que 1!=1 y 0!=1
8) Cálculo del número combinatorio o coeficiente binomial , n sobre r, que desarrolla los coeficientes del binomio de Newton

$\binom{n}{r} = \frac{n!}{r!(n-r)}$, dónde r va tomando sucesivamente los valores 0 , 1 , 2 , 3, ..., n

$$\binom{n}{0} = 1 \; ; \; \binom{n}{1} = 1$$

---ALGUNOS DE LOS SIMBOLOS EMPLEADOS

-operadores matemáticos

+ suma (a+b) ; **-** resta (a-b) ; ***** , **.** , **nada** : multiplicación (a*b , a.b , ab) ;

: , **/** , **—** división (a:b , a/b , $\frac{a}{b}$) ; **±** más-menos (sumar y restar) ; **=** igual ;

≈ aproximadamente igual ; **<** menor ; **>** mayor ; **≤** igual o menor ;
≥ igual o mayor ; **≠** , **<>** **(< >)** no igual, distinto
|a| valor absoluto de a, siempre positivo ; ΣX^2 suma de todos los cuadrados de X ;
$(\Sigma X)^2$ el cuadrado de la suma de todas las X.

-otros

Δ incremento ; **α** letra griega alfa ; **β** letra griega beta ; **λ** letra griega lambda ; **r** coeficiente de correlación ; **€(a÷b)** intervalo que va desde a hasta b ; **Σ** sumatorio abreviado, que para simplificar es el único que utilizaremos. El símbolo normal es $\sum_{i=1}^{i=n} x_i$, que quiere decir sumar todos los valores de x, desde el primero hasta el que ocupa el lugar n . si la variable x vale 10 , 12 y 14 , ΣX=36

> Clásicamente se utilizan letras griegas para simbolizar parámetros de poblaciones y letras latinas para las muestras. Aquí se utilizarán en aras de la sencillez siempre letras latinas tanto para poblaciones como para muestras, poniendo en caso de que pueda haber duda o confusión el subíndice p o m.

---LECTURA DE FORMULAS

consiste en traducirlas al lenguaje gramatical y lógico, separándolas en sus distintas partes, lo que nos permitirá resolverlas.

$F = \sqrt{\frac{\Sigma(x-5)^2}{2}}$ quiere decir: a cada valor de la variable x le restamos 5 y esta diferencia la elevamos al cuadrado; luego sumamos todos los resultados obtenidos; esta suma se divide por 2 ; finalmente se extrae la raíz cuadrada del cociente. Así obtenemos el valor de F. No hay que asustarse de fórmulas muy complejas que se resuelven de forma similar, por partes. Como dice un proverbio indio: es posible comerse todo un elefante siempre que sea a trocitos...

---RESOLUCION DE LOS CALCULOS ESTADISTICOS

Muchos se pueden resolver manualmente, utilizando lápiz , papel y los conocimientos adecuados, facilitando el trabajo las calculadoras de bolsillo. Con una calculadora científica sencilla se pueden resolver todos los problemas de esta asignatura. Es absolutamente necesario estar familiarizado con el manejo del aparato para evitar errores. Existen programas estadísticos para ordenadores, algunos gratuitos, que se verán en los temas 21 y 22 . La hoja de cálculo **Excel** permite resolver muchos problemas. En todo caso, si no se sabe Estadística, el ordenador y los programas sirven de muy poco.

---REDONDEO DE NUMEROS

Redondear un número es expresarlo por otro más corto, con menos cifras; en general comporta una pequeña pérdida de exactitud. El redondeo puede hacerse voluntariamente para obtener números más manejables o más fácilmente comprensibles. En otros casos el redondeo es obligado, como en el caso de tener que expresar un número con la sensibilidad que le corresponde (cifras significativas). Cualquier número puede redondearse, pero sobre todo se aplica a números con muchas cifras, poco frecuentes en Estadística, o con decimales. En este último caso el redondeo se indica diciendo el nº de decimales deseado o bien el lugar del redondeo (décimas, centésimas, milésimas...).

Regla general del redondeo: se redondea al número más próximo. Siempre hay dos opciones, una por encima y otra por debajo del número original.

Ejemplos: 4,1 redondeado a enteros es 4 (hay que elegir entre 4 y 5; el 4 está más cerca).
25,8 redondeado a enteros es 26 , que es el número más próximo entre 25 y 26
3,1785 redondeado a 2 decimales es 3,18 (se elige entre 3,17 y 3,18)
3,141592 redondeado a todos los lugares posibles::

redondear a	elección entre		nº redondeado
unidades	3	4	3
1 decimal	3,1	3,2	3,1
2 decimales	3,14	3,15	3,14
3 decimales	3,141	3,142	3,142
4 decimales	3,1415	3,1416	3,1416
5 decimales	3,14159	3,14160	3,14159

Caso especial del 5 como última cifra para redondear al lugar anterior : se redondea al número par.

Ejemplos: 2,5 (¿2 ó 3?) → **2** ; 2,55 (¿2,5 ó 2,6?) → **2,6** ;
2,145 (¿2,14 ó 2,15?) → **2,14** ; 2,1235 (¿2,123 ó 2,124?) → **2,124**

Más ejemplos:

5 ! = 5 * 4 * 3 * 2 * 1 = 120

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \quad ; \quad \binom{5}{3} = \frac{5!}{3! * 2!} = 10$$

$\Sigma x \quad \Sigma x^2 \quad (\Sigma x)^2$:

si x = (1 , 2 , 3 , 5) :

$\Sigma x = 11 \quad \Sigma x^2 = 39 \quad (\Sigma x)^2 = 121$

redondear 6'28945 a todos los lugares posibles:

6 6'3 6'29 6'289 6'2894

Tema 3: Variables. Medidas. Síntesis de datos estadísticos.

--Variables. Como ya se vio en el tema 1, las variables son características que se distinguen por la variabilidad con que se manifiestan en los diversos individuos.

--Tipos de variables.

Hay variables: cualitativas (CL) y cuantitativas (CT)

nombre	datos	expresión	variantes	ejemplo	
CUALITATIVAS O ATRIBUTOS	Categóricos	modalidades o categorías	2 modalidades más de 2 mod.	sexo caras dado	mujer-hombre 1, 2, 3, 4, 5, 6
CUANTITATIVAS	métricos	valores	-continuas -discretas	talla nº hijos	170 cm. 0, 1, 2, 3, ...

--Medida de las variables

Se hace según las llamadas escalas. Básicamente hay 4 escalas de medidas:

- nominales
- ordinales
- de intervalo
- de razón

Las variables ordinales son una variante de las nominales y las de razón de las de intervalo.

--Escalas nominales

Se utilizan para medir atributos, es decir, variables cualitativas. Se da un nombre a cada una de las modalidades, se asignan los individuos a ellas y se cuentan los individuos de cada modalidad (frecuencia). El orden en que se designan las modalidades es indiferente, p.e. alto y bajo o bajo y alto.

Ejemplo: la variable sexo tiene dos modalidades, hombre y mujer. Medimos este atributo en 100 personas y encontramos 52 hombres y 48 mujeres.

En vez de dar un nombre convencional a las modalidades se las puede designar con un número, lo que facilita sobre todo el tratamiento informático. Estos números son realmente un nombre y por tanto no pueden hacerse con ellos operaciones matemáticas. Así podríamos llamar a los hombres “1” y a las mujeres “2” (ó 7 y 8...)

--Escalas ordinales

Una escala ordinal es una escala nominal en la que las diversa modalidades guardan entre sí una relación de orden o jerarquía, que debe ser respetada, siendo indiferente que el orden sea de mayor a menor o viceversa. Ese orden viene marcado por el sentido común y también por la costumbre.

Un ejemplo clásico son las notas académicas tradicionales : sobresaliente-notable-aprobado-suspenso o suspenso-aprobado-notable-sobresaliente. En la variable “evolución de la enfermedad” podríamos distinguir las siguientes modalidades : muerto-peor-igual-mejor-curado , o bien, curado-mejor-igual-peor-muerto.

También pueden emplearse números como nombre de modalidades, pero respetando el orden. Podríamos hacer muerto=1, peor=2, igual=3, mejor=4, curado=5 . O bien, curado=1, mejor=2, igual=3, peor=4 , muerto=5 .

--Escalas de intervalo

Se utilizan para medir variables cuantitativas cuando no hay cero absoluto en la zona de medición, lo que permite valores negativos. El cero se asigna arbitrariamente así como la unidad de medida.. La escala ha sido diseñada de tal manera que sus números permiten valorar exactamente la diferencia que hay entre dos medidas (= intervalo). Ejemplo típico es la temperatura medida de la forma habitual, lo que puede hacerse de diversas maneras. En Europa se mide en grados

centígrados o Celsius (C). El " 0" se asigna a la temperatura de congelación del agua destilada y el "100" a su temperatura de ebullición. Ese intervalo se divide en 100 partes y así se obtienen los grados centígrados. En USA se mide en grados Fahrenheit (F). 0º C equivalen a 32º F y 0º F equivalen a –17,78º C. Por tanto 32º C no representa el doble de calor que 16º C, simplemente el doble de grados C. Esas temperaturas medidas en grados Fahrenheit serían 0º F y –8,9º F. Un niño con un proceso febril en Castellón puede tener 40º C de fiebre; en USA tendría 104º F. Por la Física sabemos que hay un mínimo infranqueable de temperatura, el llamado "cero absoluto", que en grados centígrados corresponde a –273,15º. Este cero no significa la ausencia de temperatura, sino el mínimo de temperatura posible. La escala de Kelvin asigna su 0 a esta temperatura.
--Escalas de razón
Se utilizan para medir variables cuantitativas cuando hay un cero absoluto, siendo la unidad de medida lo único arbitrario. Una longitud puede ser medida en cm., Km., yardas, varas, etc. pero el cero es el mismo para todos. El tiempo de reacción a un estímulo siempre empieza en cero cualquiera que sea el sistema que utilicemos para medir el tiempo. Aquí sí puede decirse que una persona que pesa 50 Kg. pesa el doble que un niño que pesa 25. Y que la diferencia de peso entre una persona que pese 80 Kg. y otra que pese 50 Kg. es la misma que la existente entre dos piedras de 35 y 5 Kg., respectivamente. No hay valores negativos.

--Variables cualitativas
Las variables cualitativas (CL) o atributos se miden por escalas nominales u ordinales según corresponda. Cuando sólo tienen dos modalidades se llaman dicotómicas. Ejemplos: cara-cruz, varón-hembra, vivo-muerto. Todos los atributos, con independencia del número de modalidades que tengan, pueden ser siempre reducidos a dicotómicos si así se desea. Los 4 palos de la baraja española (oros, copas, espadas y bastos) pueden ser reducidos a oros-no oros, bastos-no bastos, etc. ; las marcas de coches a Seat-no Seat. ; el estado civil a casado-no casado...

--Variables cuantitativas
Las variables cuantitativas (CT) se miden por escalas de intervalo o de razón, según su naturaleza. Pueden ser continuas o discretas.
Una variable CT es continua cuando puede tomar cualquier valor en su zona de variabilidad. Son continuas la talla, el peso, la tensión arterial, el contenido de un frasco, la glucemia, etc.
Las variables CT discretas no pueden adoptar cualquier valor, sino solamente ciertos valores. Una familia puede tener 0, 1, 2, 3, ... hijos, pero no 3,1416 hijos. El nº de pacientes que ingresa en un hospital,, el nº de ataques que sufre un paciente en un mes, el nº de cápsulas de un envase medicamentoso... son discretas.
Una variable CT continua se mide a menudo, porque resulta más práctico, de forma "discretizada" . La edad suele expresarse en años enteros, o en meses en los niños pequeños, pero no por eso deja de ser continua.

--Transformación de variables
Las variables cuantitativas pueden ser transformadas en cualitativas, dicotómicas o no, con una pérdida en la calidad de la medida, que a veces se asume si mejora la información. La talla podemos medirla en alta-normal-baja. Los valores de colesterol en mayor de 200 mg/dl - igual o menor de 200 mg/dl. Como la variable CT proporciona más información que la CL debe ser usada siempre que no sea más conveniente hacerlo de forma cualitativa.

Las variables CL en cambio no pueden ser transformadas en CT.

Las variables CL son por su propia naturaleza discretas. Por las limitaciones de los instrumentos de medida la mayoría de las CT continuas son discretizadas.

Dos ejemplos:
---variable "INGESTION DE ALCOHOL" .
He seleccionado 4 formas distintas en orden creciente de información:

1) abstemio – bebedor — Variable CL con dos modalidades, nominal.
2) abstemio – bebedor – alcohólico — Variable CL con tres modalidades, ordinal.
3) nº de copas o vasos bebidos en una semana — Variable CT discreta
4) gramos de alcohol tomados en una semana — Variable CT continua

---"ESTUDIO DE 3 TRATAMIENTOS DE LA ISQUEMIA CORONARIA".
Considerando las variables:

- sexo : hombre – mujer — CL con 2 modalidades, nominal
- medicamento: A – B – C — CL con 3 modalidades, nominal
- nº ataques del día anterior — CT discreta
- distancia caminada sin disnea — CT continua

--Necesidad de una definición clara de las variables

Es esencial que todo el mundo sepa qué se está midiendo y cómo. Está claro lo que es medir el peso en Kg. o la talla en cm. Pero, ¿que es ser "fumador"?. ¿El que fuma un pitillo, aunque sea una vez al año? ¿O el que fuma cada día o al menos cada tres?... Hay que concretar y decir por ejemplo: "en este estudio se considera fumador a quien fuma al menos un cigarrillo cada semana" o "se considera desnutridos a los niños que en los gráficos peso/talla de Tanner están por debajo del percentil 3", etc., etc. ....

--Dominio de una variable

Es el conjunto de valores o modalidades que puede adoptar. El dominio de la variable CL "puntuación de la cara de un dado" es (1, 2, 3, 4, 5 y 6). El de la variable sexo: (hombre, mujer). El de la "longitud de las hojas de la planta P" cualquier valor entre1 y 8 cm. o $\in (1 \div 8)$, etc.

--Variables aleatorias y controladas

Una variable es <u>controlada o independiente</u> cuando su valor o la modalidad elegida en cada uno de los individuos depende únicamente del investigador. En un estudio podemos seleccionar sólo individuos del sexo masculino. O fijar la dosis de medicamento que se da a los ratoncillos, etc.
Una variable es <u>aleatoria o dependiente</u> cuando su valor en cada uno de los individuos no depende del investigador, sino de la naturaleza o reacción del propio individuo. Por ejemplo la talla de los alumnos de una clase, la tensión arterial de un grupo de pacientes, etc.

--Medida de una variable continua

Debido a la imperfección de los instrumentos de medida, aún de los más sofisticados, el <u>valor exacto o real</u> de una medida (**Xe**) es realmente desconocido y sólo podemos expresarlo de una forma aproximada mediante <u>el valor medido</u> (**X**). Supongamos que estamos midiendo una longitud con una regla graduada. Cuando la medida no se corresponde con un valor marcado en la regla, hay que aproximar (=redondear) a la marca más cercana. Si hay equidistancia se aproxima al valor par.

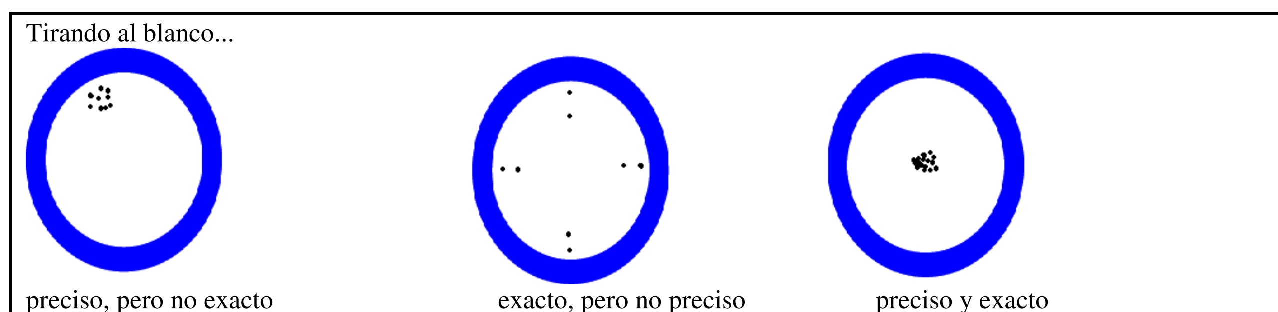
```
            5      6      7       8
            | -x - -- | - - -x - | - -x - - |
medida:       5          7      8
```

La diferencia entre el valor exacto y el valor medido se llama ERROR ABSOLUTO. Toda medida tiene su error.

$$\mathbf{E} = |\mathbf{Xe} - \mathbf{X}| \quad \text{y por tanto} \quad \mathbf{Xe} = \mathbf{X} \pm \mathbf{E}$$

Este error, del que sólo podemos conocer su máximo (**Emax**), depende de la sensibilidad, precisión y exactitud de los instrumentos de medida.
La sensibilidad (se) es la unidad más pequeña que permite utilizar el instrumento de medida. En las reglas graduadas habituales se = 1 mm. El **Emax** es igual a la mitad de la sensibilidad; **Emax = se/2**. Una regla milimetrada: tiene un **Emax** de 1/2 mm. = 0,5 mm.
Hay precisión cuando repetida la medida muchas veces da valores iguales o muy parecidos.
Hay exactitud si la media de repetidas medidas coincide con el valor exacto de la medida.

Así, si una longitud real de 9,0 cm. la medimos 4 veces y obtenemos 9,1; 9,0; 9,0 y 8,9 el instrumento es preciso y exacto. Si obtenemos 5,6 ; 5,5; 5,7; 5,6 será preciso, pero no exacto. Midiendo 9 ; 6 ; 12 ; 3 y 15 será exacto pero no preciso. La medida ideal es la que se obtiene con un máximo de sensibilidad, precisión y exactitud.



--¿Que sensibilidad se debe utilizar?

Una sensibilidad escasa proporciona datos de poca confianza, con mayor margen de error. Si es excesiva no es mala en sí, pero en general supone aparatos más caros y de manejo más difícil. Hay que elegir la más adecuada para cada caso concreto, teniendo en cuenta la experiencia y el sentido común.

La sensibilidad es adecuada si la diferencia entre la medida más alta ,sin punto o coma decimal, y la medida más baja , también sin punto o coma decimal, está entre 30 y 300 .

Ejemplo:
1- medimos en mm. la longitud de las hojas de la planta XYZ. La medida mayor es 8 y la menor 4. Como 8-4=4 , que es menor de 30, la sensibilidad utilizada no es buena. En una medida de 5 mm. el error máximo es de 0,5 mm., o sea de un 10%. El instrumento de medida no es adecuado.
2- después utilizamos un aparato que mide en décimas de mm. Como valores extremos obtenemos 8,4 y 4,3 mm. 84-43=41, que está entre 30 y 300. En una medida de 5,0 mm. el error máximo es de 0,05 mm., un 1%. Este instrumento sí es adecuado.

--Valor puntual y por intervalo de una medida

Al desconocer el valor exacto de una medida, **Xe ,** hay que estimarlo. La medida se puede expresar de dos formas: puntual o por intervalo.
La medida puntual o valor puntual es el valor medido, X ; por tanto no es exacto..
El valor por intervalo o medida por intervalo es el intervalo en el que con seguridad (¡si se ha medido bien!) estará el valor exacto **Xe** de la medida. Se obtiene sumando y restando al valor puntual el error máximo, es decir, la mitad de la sensibilidad: $X \pm se/2$. De esta forma se obtienen los llamados *límites reales de la medida*, uno superior y otro inferior. Si medimos un lápiz con una regla milimetrada y obtenemos 151 mm., la medida puntual será 151 mm. Como la sensibilidad es de 1 mm., la medida por intervalo será 151±0,5 mm. o $\in$(150,5÷151,5).
Si utilizamos una regla con nonius, que mide en décimas de mm. y obtenemos 151,1 mm. , el valor puntual será 151,1mm. Aquí la sensibilidad es de 0,1 y por tanto la medida por intervalo será 151,1±0,05 ó $\in$(151,05÷151,15).

Como es fácil equivocarse al realizar los cálculos, puede resultar útil el procedimiento siguiente:
a) se toma el número, prescindiendo del posible punto o coma decimal y se añade un 0
b) se le suma y resta 5
c) si había decimales, se vuelve a poner la coma o punto decimal en su sitio. Así tenemos los dos límites del intervalo.
En el último ejemplo: 151,1 → 15110 → *-5 = 15105 y +5= 15115* → 151,05 y 151,15

--Cifras significativas

Son las cifras del valor puntual de una medida, prescindiendo de los ceros a la izquierda de la primera cifra con valor distinto de cero. Son pues función de la sensibilidad.

medida	cifras significativas		medida	cifras significativas
65,5 m	3		4,53400 cm	6
0,0018 kg	2		1,00180 amp	6
1,0018 mm	5		0,10000 sec	5

En un número redondeado las cifras significativas llegan tan sólo hasta el lugar del redondeo. 18 millones como redondeo de 18 234 156 tiene 2 cifras significativas ; 3,14 como redondeo de 3,141592 tiene 3.

--Métodos de recuento (variables CL)

a) observación, utilizando los órganos de los sentidos.
b) gráficos: métodos de palotes, cuadrados...
c) tarjetas de formas, contenidos o colores distintos
d) lectura óptica, como en el escrutinio de quinielas y similares
e) lectura magnética (de espacios marcados con lápiz de grafito)

--Síntesis de datos estadísticos

Una vez medida la variable en los diversos individuos se tiene una serie de datos, métricos o categóricos, los llamados DATOS ORIGINALES o DATOS AISLADOS, que sin más elaboración suelen ser poco útiles.
Es necesario ordenarlos y resumirlos para que proporcionen la máxima información de la forma más sencilla posible. Esto se hace de diversas formas:

- agrupando los datos según su frecuencia, con lo que se transforman en DATOS AGRUPADOS O DISTRIBUCION DE FRECUENCIAS, construyendo las correspondientes TABLAS y GRAFICOS ESTADISTICOS
- calculando los llamados INDICES o PARAMETROS ESTADISTICOS, como media aritmética, desviación estándar, porcentajes, etc.

Las Escuelas clásicas utilizan el término INDICE para las muestras y sus símbolos se representan con letras latinas, mientras que el término PARAMETRO se reserva para las poblaciones, con símbolos de letras griegas. Aquí utilizaremos ambos términos de forma indistinta, es decir, tanto para poblaciones como para muestras. Y salvo alguna rara excepción los símbolos serán de letras latinas.

Recordatorio : MEDIDA DE UNA VARIABLE CONTINUA

Xe	valor exacto, real, de la medida ; es desconocido
X	valor medido por el instrumento; es el valor puntual
E = |Xe – X|	error de la medida ; por tanto Xe = X ± E
E Máximo (**Emax**)	se/2
Valor por intervalo de una medida	X ± Emax ó ∈(X-Emax ÷ X+Emax) en ese intervalo está contenido el valor real Xe

Tema 4 : Tabulación de datos

La tabulación consiste en presentar los datos estadísticos en forma de tablas o cuadros.

--Partes de una tabla

- TITULO de la tabla, que debe ser preciso y conciso
- CONTENIDO, con
  - la *fila de encabezamiento o cabecera* (títulos de las columnas)
  - la *columna matriz*, con las modalidades o clases de la variable
  - *columnas de parámetros*
- NOTAS EXPLICATIVAS (opcional), como fuente de los datos, abreviaturas, etc.

--Forma de tabular

VARIABLES CUALITATIVAS

pueden representarse :

- la frecuencia absoluta (símbolo : **f** ó **n**), que es el nº de veces que aparece cada modalidad (resultado del recuento). La frecuencia total, de todas las modalidades juntas, se representa por **N**.
- la frecuencia relativa (**fr**) o proporción se obtiene dividiendo la frecuencia de cada modalidad entre el total de datos. fr = f / N . Los valores posibles oscilan entre 0 y 1. Suele expresarse con 3 decimales. La suma de todas las fr tiene que dar 1 ó un número muy cercano al 1, si ha habido redondeos.
- el porcentaje (**P** o **%**), que es la frecuencia relativa multiplicada por 100. P = fr * 100 ó % = (f*100)/N . Suele expresarse con 3 dígitos. La suma de todos los porcentajes debe dar 100 o un número muy próximo, si ha habido redondeos.
- las frecuencia acumuladas (**Σ f** ó **Σn**) que se obtienen sumando la frecuencia de cada modalidad a las frecuencias ya acumuladas anteriormente. En la primera modalidad no hay nada acumulado de antes y por tanto su frecuencia acumulada será su misma frecuencia. La última modalidad tiene que dar una frecuencia acumulada igual a N.
- las frecuencias relativas acumuladas y los porcentajes acumulados se obtienen de forma similar
- En las variables nominales las modalidades pueden ponerse en el orden que se quiera, pero en las ordinales hay que respetar el orden lógico.

Ejemplo:

Residencia Sanitaria S. S. de Castellón
Ingresos en Pediatría. Marzo 1980

Sección	f	fr	%	Σf	Σfr	Σ%
Neonatología	25	0,125	12,5	25	0,125	12,5
Lactantes	95	0,475	47,5	120	0,6	60
Preescolares	80	0,400	40	200	1	100
Total	200	1	100			

En la tabla definitiva no se presentan todos estos parámetros, sino los más adecuados en cada caso concreto. Casi siempre f y/o % . Sólo el porcentaje, sin que conste N, no es correcto. En este ejemplo bastaría con f y % .

VARIABLES CUANTITATIVAS

Los datos se agrupan según la frecuencia de los valores. Es lo que se denomina *Distribución de frecuencias*. La forma de tabular depende del nº de datos.

----Si son pocos (la mayoría de autores pone el tope en 30) , se hace una tabla simple de forma similar a lo visto para las variables CL. Cada dato equivale a una modalidad. Al final nos quedaremos con la f de cada número y si se prefiere también con el %. Los números se ordenan de menor a mayor o de mayor a menor. La tabla puede hacerse en sentido vertical u horizontal.
Ejemplo: Si x = (4 , 1 , 7 , 2 , 2 , 9 , 7 , 2 , 2 , 9 , 7 , 1 , 4)

x	1	2	4	7	9
f	2	4	2	3	2

o bien

x	f
1	2
2	4
4	2
7	3
9	2

----Si son muchos se agrupan en clases, que son intervalos sucesivos de valores. Los datos se asignan a la clase que les corresponde y se cuentan los datos de cada clase, que está representada por el punto medio o centro de clase (pm ó c).

Esta agrupación es **arbitraria** con dos condiciones esenciales: que las clases sean mutuamente excluyentes y que todos los datos puedan se asignados a una clase. Ahora bien, la experiencia ha ido introduciendo una serie de normas, que permiten hacer esta agrupación de la forma más racional posible.

Yo recomendaría los siguientes pasos:

1) calcular el **RECORRIDO (R) ,** (a veces mal llamado Rango)
= (límite real superior del dato mayor – límite real inferior del dato menor)
O si se prefiere: = (valor tabulado máximo – valor tabulado mínimo) + 1
2) calcular el **Nº DE CLASES** (NC) .
Es función de N (tamaño de la muestra) y no hay reglas fijas.
En general: “entre 4 y 20” .
Ayudas: NC = 1+ 3,32*logN ó 1+1,44*lnN
O la siguiente tabla:

N	8	16	32	64	128	256	etc.
NC	4	5	6	7	8	9	etc.

De entrada nos quedamos con 2 ó 3 opciones

3) calcular la A**MPLITUD** de las clases ó INTERVALO (i) : i = R / NC
Si **i** no es número entero, se redondea al número entero superior para que NC*i ≥ R y así queden englobados todos los datos
Como probamos con 2 ó 3 opciones, conviene elegir una i que sea impar, pues así el punto medio de la clase (pm ó c) tendrá una cifra menos.

En principio todas las clases deben tener la misma amplitud.

4) Ver si hay **SOBRAS**, que son la diferencia entre NC*i y R. Se reparten lo mejor posible entre ambos extremos de la distribución fijando así los límites definitivos de la tabla.

5) Construir el esquema de la tabla, poniendo **columnas** de

- CLASES ó LIMITES TABULADOS
- LIMITES REALES
- PUNTO MEDIO (pm ó c)
- FRECUENCIA (f ó n)
- FRECUENCIA RELATIVA (fr)
- PORCENTAJE (P o %)
- FRECUENCIAS ACUMULADAS (Σf ó Σn)
- FRECUENCIAS RELATIVAS ACUMULADAS (Σfr)
- PORCENTAJES ACUMULADOS (Σ%)

6) Hacer el **RECUENTO** de datos y rellenar las casillas correspondientes

7) Escribir la **TABLA DEFINITIVA.** Son obligadas las clases y la frecuencia absoluta, pudiendo añadir otros parámetros, si se considera que mejoran la información. Una tabla excesivamente prolija resulta más difícil de leer. Por tanto la norma es: poner todo lo necesario, pero no más de lo necesario.

Es recomendable probar con al menos 2 tablas y elegir la que quede mejor.

Algunos de éstos parámetros son los mismos que se han visto para las variables CL. Otros precisan una aclaración:

Los **<u>límites de las clases</u>** son los valores inferior y superior de cada clase. (Límite inferior y límite superior). Hay que distinguir entre los <u>límites tabulados (LT)</u> y los <u>límites reales (LR)</u>. Los límites tabulados son los datos originales que abren y cierran una clase. Los límites reales son el límite real inferior del primer valor (LRI) y el límite real superior del último (LRS).

El **<u>punto medio o centro</u>** de la clase (pm ó c) representa a la clase cuando se hacen operaciones matemáticas. Es la media de los límites. Da lo mismo tomar los límites reales que los tabulados, ya que ambos dan el mismo resultado.

En una distribución con todas las clases de la misma amplitud las diferencias entre los puntos medios, los límites inferiores y los límites superiores de dos clases consecutivas valen lo mismo y son igual a la amplitud de la clase (i). Esto facilita la construcción de la tabla.

Una **clase** es **abierta** cuando carece de un límite. Sólo pueden ser abiertas la primera clase (p.e. <10 ; no tiene límite inferior)) y la última (p.e. >100 ; no tiene límite superior). No deben usarse, a no ser que no haya otro remedio.

EJEMPLO:

Tabular los 70 valores siguientes:

DATOS ORIGINALES (N = 70)

40 55 19 51 62 15 20 44 60 60 45 15 21 31 13 44 41 43 51 35 50 33 25 16 61 14 14 59 59 59 20 23 25 29 29 59 58 54 50 49 39 27 37 23 24 58 27 28 57 32 32 34 57 56 35 35 54 36 43 46 52 50 49 42 43 46 40 39 31 48

PASOS DE LA TABULACION

-dato mayor: 62, cuyo LRS es 62,5
-dato menor: 13, cuyo LRI es 12,5
-recorrido (R): 62,5-12,5 = 50 ó (62-13)+1 = 50
-nº de clases (NC): 7 u 8
-amplitud (i):
- -si NC = 7 , i = 50/7 = 7,1 → 8 (par)
- -si NC = 8 , i = 50/8 = 6,2 → 7 (impar)

-nos quedamos pues con NC = 8 de amplitud 7, que es impar
-sobras: (8*7) – 50 = 6 , que repartimos así: 3 abajo y 3 arriba
- la 1ª clase empezará en 10 (13-3)
- la última terminará con el 65 (62+3)

--ya se puede construir el esquema de la tabla (clases, LR y punto medio) y proceder al recuento de los datos que corresponden a cada clase, para completar las otras columnas

Clases (Límites tabulados)	Límites reales	punto medio c	f	fr	%	Σf	Σfr	Σ%
10 – 16	9,5 – 16,5	13	6	0,09	8,57	6	0,09	8,57
17 – 23	16,5 – 23,5	20	6	0,09	8,57	12	0,17	17,1
24 – 30	23,5 – 30,5	27	8	0,11	11,4	20	0,29	28,6
31 – 37	30,5 – 37,5	34	11	0,16	15,7	31	0,44	44,3
38 – 44	37,5 – 44,5	41	11	0,16	15,7	42	0,60	60,0
45 – 51	44,5 – 51,5	48	11	0,16	15,7	53	0,76	75,7
52 – 58	51,5 – 58,5	55	9	0,13	12,9	62	0,89	88,6
59 - 65	58,5 – 65,5	62	8	0,11	11,4	70	1,00	100
Suma			70	1,01	99,94			

***Esta no es la única tabla posible, aunque probablemente sea la mejor.
Podríamos hacerla con 7 clases de amplitud 8; sobras: 6 . Clases: 10 – 17 ; 18 – 25; ...; 58 - 65
O bien 6 clases de amplitud 9. Sobras: 4 . Clases: 11- 19 ; 20 – 28; ...; 56 - 64
o bien 10 clases de amplitud 5 . Sin sobras. Clases: 13 –22 ; 23 – 32 ; .... ; 53 - 62

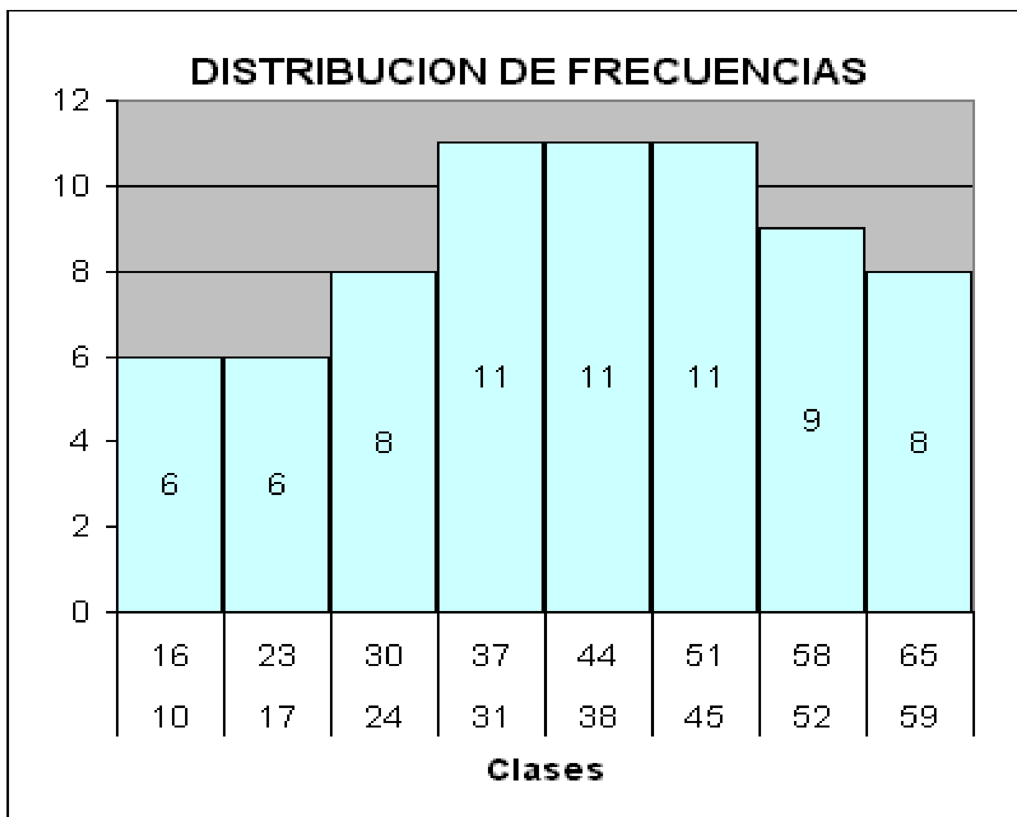
***En la tabla definitiva no suelen ponerse los LR. Las clases y la frecuencia están prácticamente siempre. Según la naturaleza de la variable puede ser conveniente añadir algún otro parámetro, que contribuya a una información mejor y más clara.

*** En la página siguiente puede verse la tabla y el gráfico que elabora automáticamente mi programa de Excel, Exceltabla.xls, a partir de los 70 datos del ejemplo anterior, introducidos en la columna A.

Tabla e histograma del ejemplo de la página 4-4 que hace "Exceltabla"

Lim.Tab.Inf.	LimTab.Sup.	pm	f	%	Σ f	Σ %
10	16	13	6	8,6	6	8,6
17	23	20	6	8,6	12	17,1
24	30	27	8	11,4	20	28,6
31	37	34	11	15,7	31	44,3
38	44	41	11	15,7	42	60,0
45	51	48	11	15,7	53	75,7
52	58	55	9	12,9	62	88,6
59	65	62	8	11,4	70	100,0
			0			

Datos origin.:	SESGO	-0,196	MODA	59,00
	CURTOSIS	-1,105	p3	14,07
	MEDIA GEO	36,53	p10	19,90
	MEDIANA	40,50	p25	28,25
	MEDIA	39,59	p75	51,00
	DS	14,41	p90	59,00
	VARIANZA	207,58	p97	60,00



Tema 5 : Representaciones gráficas

Los datos estadísticos pueden ser también representados por medio de **gráficos**. Un viejo proverbio chino dice que una imagen vale más que mil palabras (o que mil números, aplicado a la Estadística). Los gráficos son una simplificación y un complemento de una tabla estadística. Son más sencillos, más llamativos y a menudo más inteligibles, aunque se pierde información.

Componentes

Como en las tablas estadísticas se pueden distinguir:

- el título
- el gráfico en sí (casi siempre complementado con números)
- notas explicativas , si procede

Tipos de gráficos

-Diagramas
  - -de barras
  - -histogramas
  - -polígonos de frecuencias

-Gráficos sectoriales
-Pictogramas
-Otros

Los **DIAGRAMAS** utilizan un sistema de coordenadas cartesianas. En el eje de abscisas (*x*) se representa la variable. En el de ordenadas (*y*) las frecuencias o porcentajes.
Si la variable es CL se marcan en el eje de abscisas las modalidades y sobre ellas se dibujan líneas o barras de altura proporcional al parámetro representado. Si la variable es CT se marcan los valores y clases correspondientes al recorrido de la variable.
La escala de *y* debe de empezar siempre en 0 para evitar manipulaciones y engaños ópticos. Habitualmente se trata de una escala aritmética, pero cuando hay frecuencias o valores muy dispares el gráfico es apenas legible y es mejor utilizar escalas logarítmicas o semilogarítmicas. Una alternativa, algo chapucera, es quebrar claramente la escala y las barras. Todo antes que violar la norma del comienzo de *y* en 0.
En un buen diagrama la longitud de x debe de estar entre 1 y 2 veces la de y. Ambas escalas deben de estar claramente rotuladas, directamente o por medio de una nota explicativa. Son preferibles números cortos (redondeados) y hay que evitar dar excesivos datos, sobre todo en presentaciones, ya que el gráfico se muestra un corto espacio de tiempo. Otra cosa es un gráfico impreso al que el lector puede dedicarle el tiempo que quiera. Los ordenadores permiten fácilmente dibujar los gráficos en 3D. Las barras pasan a ser prismas o incluso cilindros o conos, a gusto del usuario.

-El diagrama de barras o columnas es propio de variables discretas (todas las CL y las CT discretas). Cada barra corresponde a una modalidad o valor de la variable.. La altura de la barra es proporcional a la frecuencia a representar. Todas las barras deben de tener la misma anchura y la distancia entre ellas debe de ser como máximo la anchura de las barras.
Se pueden distinguir tres tipos de diagramas de barras:

a) simples (figuras 1 y 2)
b) de barras adosadas o parcialmente superpuestas, cuando se presentan de forma paralela dos conceptos que interesa comparar, p.e. hombres y mujeres (figuras 3 y 4)
c) de barras mixtas, apiladas, una variante del anterior (figura 5).

-El histograma es propio de variables CT continuas agrupadas en clases. Las barras están unas al lado de otras sin separación, a no ser que alguna clase tenga una frecuencia de 0. Cada barra

empieza en el límite real inferior de la clase que representa y termina en el límite superior, que a su vez es el comienzo de la clase siguiente. El punto medio de la clase coincide con el centro de la base. La superficie de cada barra es proporcional a la frecuencia de la clase. Si todas las clases tienen la misma amplitud, como en principio debe ser, la altura es la frecuencia de la clase. Si hay clases con distinta amplitud no puede ponerse la etiqueta de frecuencia (f ó n) en el eje vertical, ya que sería engañoso. Debe figurar la de "densidad de frecuencias" (fd). $\mathbf{fd = \frac{f}{i}}$ (fig. 6)

<u>Se pueden distinguir tres tipos de histogramas</u>:
1) <u>el H. simple</u>, que es el que acabamos de ver (fig. 7)
2) <u>el H. de frecuencias acumuladas</u>, en el que cada barra representa las frecuencias acumuladas en cada clase. El gráfico tiene forma de escalera más o menos irregular. (fig 8)
3) el <u>H. doble</u>, cuyo paradigma es la **pirámide de población**. Este gráfico nos informa de la distribución por edades de un grupo poblacional, separando hombres y mujeres y rotando el gráfico de tal forma que las edades de las personas, agrupadas en clases, están en el eje vertical y la frecuencia de cada clase en el eje horizontal. (fig. 9).

Un **POLIGONO DE FRECUENCIAS** se obtiene uniendo los puntos medios de los techos de un hipotético histograma, que se corresponden, al ser la barra un rectángulo, con los puntos medios o centros de cada clase. La línea debe comenzar y terminar en el eje de abscisas, precisamente en el sitio que correspondería al punto medio de dos clases inexistentes, la que precedería a la primera y la que seguiría a la última. Si se superponen un histograma y el correspondiente polígono de frecuencias se ve que la superficie del histograma y el área que incluye el polígono es la misma. Por tanto ambos representan igualmente a la distribución. Los hay también simples y de frecuencias acumuladas. (fig. 10 y 11)

Cuando no se representa toda la distribución sino tan sólo una parte de la misma, no hay que bajar la línea hasta el eje de abscisas. Por delante y detrás de lo representado hay clases cuya frecuencia no es ofrecida al lector. Este gráfico se llama **diagrama lineal**.

Los GRAFICOS SECTORIALES o de TARTA equivalen a un diagrama de barras y por tanto sirven para representar variables discretas. Se utilizan círculos o semicírculos y a cada modalidad o valor se le adjudica un sector circular, cuya superficie sea proporcional a la frecuencia relativa o porcentaje. Para ello se calcula el ángulo que le corresponde mediante una simple regla de tres. A todo el círculo le corresponden 360º y si es un semicírculo 180º.

En el ejemplo de los ingresos en Pediatría:
al 100% (todos) le corresponden 360º
al 12,5% (Neonatos) " " xº xº = 45º
y así para las otras Secciones se obtiene: Lactantes 171ª y Preescolares 144º

Luego mediante un transportador se trazan en el círculo las líneas correspondientes.
Los sectores circulares se pueden desgajar del conjunto para que resalten más. (fig. 12 y 13)

Los **PICTOGRAMAS** utilizan figuras e imágenes de todo tipo, como pilas de monedas, balanzas, coches, muñequitos, mapas distorsionados, etc. Siempre deben respetar el espíritu del gráfico básico. (fig. 14)

La fantasía y la inspiración pueden sugerir **OTROS** tipos de gráficos. Pero lo esencial no es que sean bonitos, sino que informen bien. Pero si son buenos, bonitos y sencillos, mejor que mejor.

Los gráficos se prestan mucho a la manipulación (no respetando las normas básicas que se han citado) y pueden ofrecer por tanto una información falsa (fig. 15 y 16). En este caso se podría decir que una imagen puede mentir más que mil palabras.

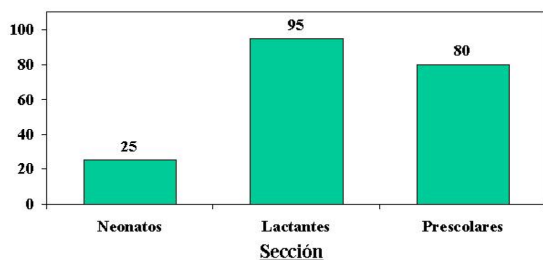


Figura 1
Diagrama de barras simple

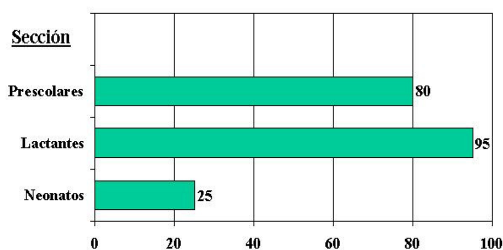


Figura 2
Diagrama de barras simple, rotado

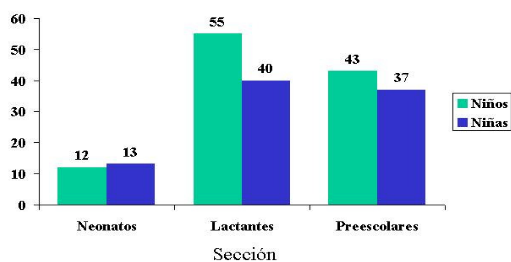


Figura 3
Diagrama de barras adosadas

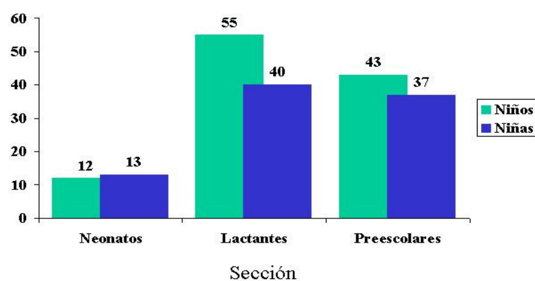


Figura 4
Diagrama de barras parcialmente superpuestas

Residencia Sanitaria de la S.S. Castellón
Ingresos en Pediatría. Marzo 1980

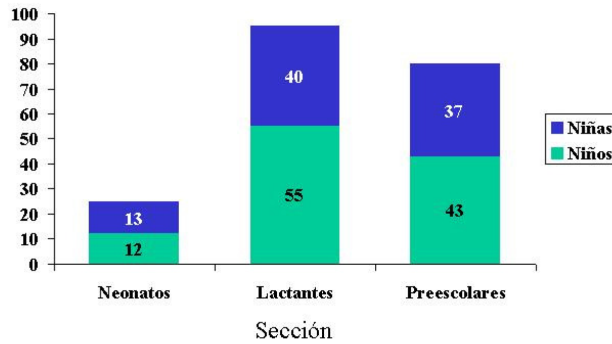


Figura 5
Diagrama de barras mixtas o apiladas

	clases	f	i	fd=f/i
A	0-3	12	4	3
B	4-8	20	5	4
C	9-11	15	3	5

La amplitud de las clases de esta distribución varía.
La superficie de las columnas representa correctamente a las clases; su altura depende no de la f sino de la df

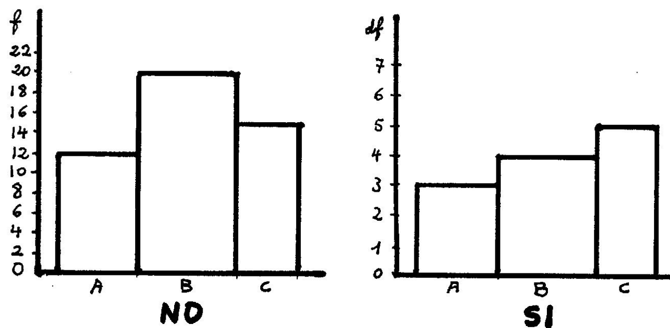


Figura 6
Si no son iguales todas las clases, hay una regla especial

"70 DATOS"
HISTOGRAMA

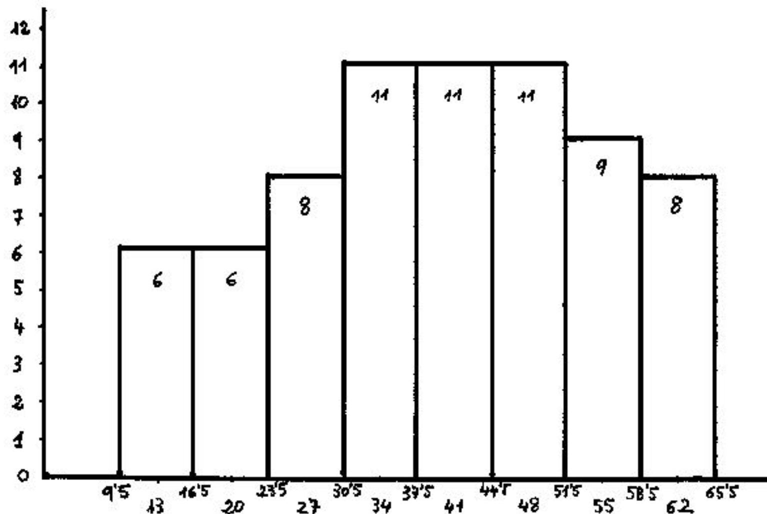


Figura 7
Histograma simple
"70 Datos" de la tabla del tema anterior

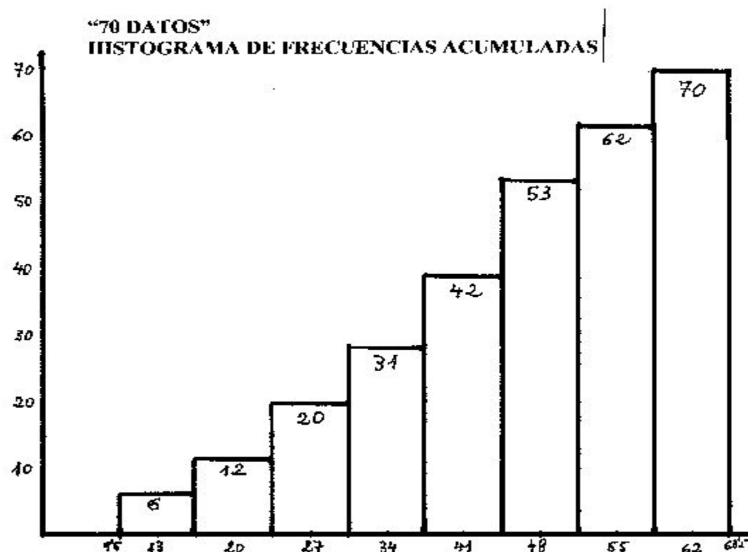


Figura 8
Histograma de frecuencias acumuladas

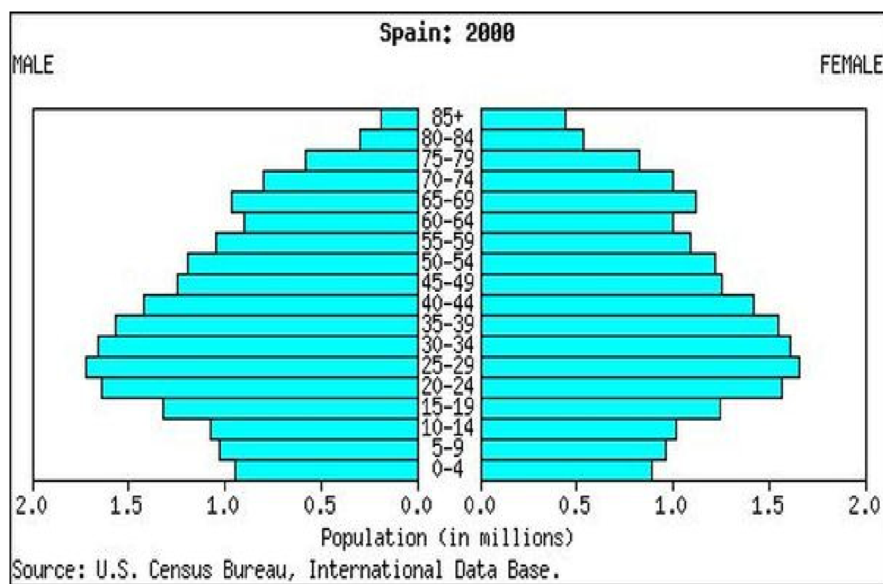


Figura 9
Pirámide de población de España en 2002.

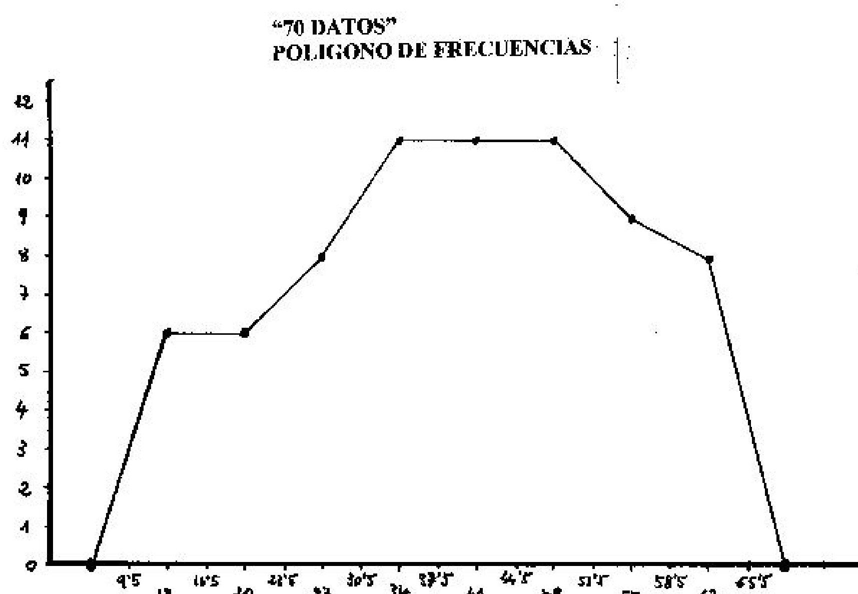


Figura 10
POLIGONO DE FRECUENCIAS

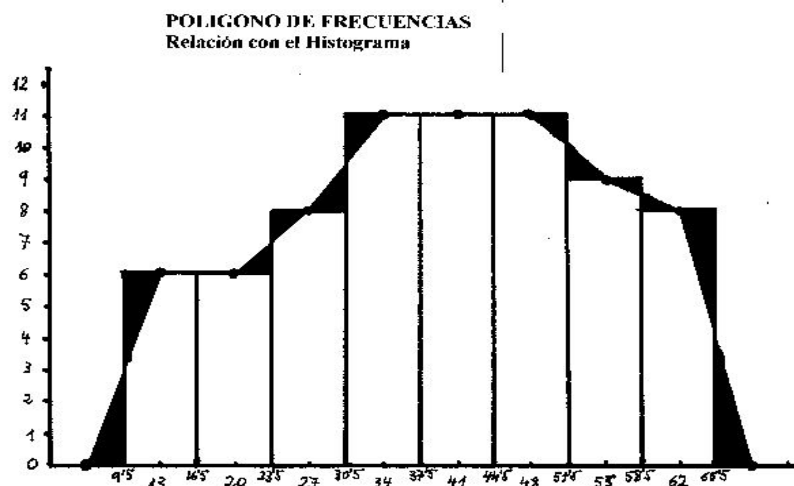


Figura 11
Relación entre el histograma y el polígono

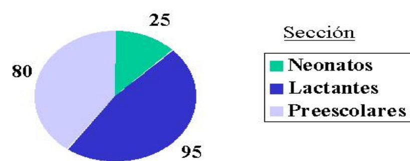


Figura 12
Diagrama circular o de tarta

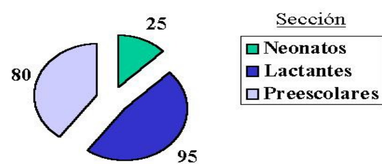
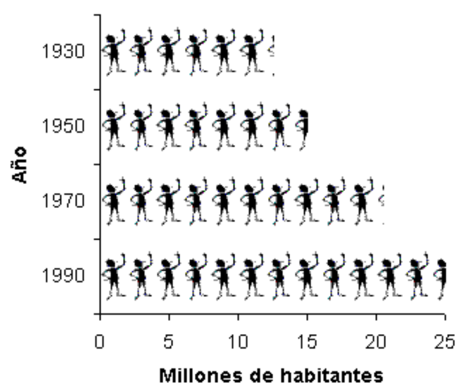
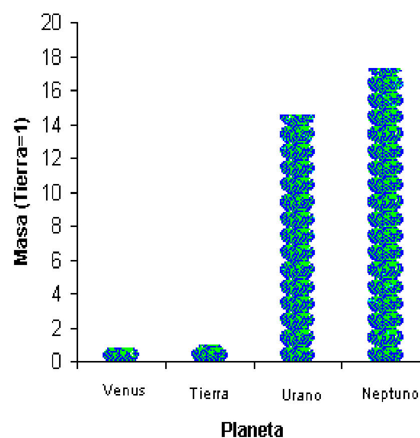
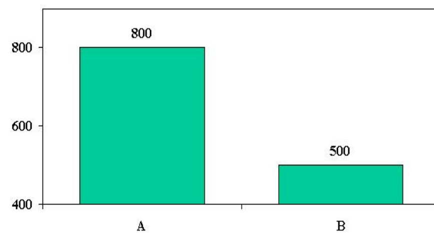


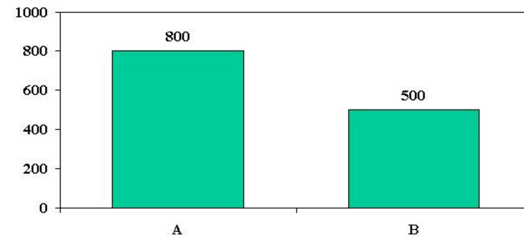
Figura 13
Diagrama circular, cortado





Figuras 14 y 15 Pictogramas





Figuras 16 y 17
El no empezar la escala en 0 , agranda las diferencias
El gráfico de la izquierda es incorrecto

Tema 6 . Índices estadísticos de variables cuantitativas. Parámetros de tendencia central, dispersión, posición y forma.

Los parámetros o índices (ya vimos en el tema 3 que consideramos ambos conceptos como equivalentes) son otra forma de presentar resumidos los datos estadísticos.
Hay que distinguir:

- parámetros de tendencia central, que informan del centro de la distribución
- parámetros de dispersión, que informan de la dispersión de los datos
- parámetros de posición, que sitúan a los datos en el conjunto ce la distribución ordenada. Los más utilizados en Bioestadística son los percentiles. Algunos de ellos pueden ser considerados también como parámetros de tendencia central y otros como de dispersión.
- parámetros de forma, que precisan la forma de la distribución. Podría decirse que expresan numéricamente la forma del histograma.

Parámetros de tendencia central

Los más importantes son:

- la media aritmética, o simplemente la media
- la mediana
- la moda
- los percentiles “centrales” (p 25 a p75)

En la explicación de los parámetros se utilizarán tres grupos de datos en los ejemplos: Supuesto A): 8 , 1 , 4 , 8 , 8 , 5 , 1 Supuesto B): los “70 DATOS” originales del tema 4 Supuesto C): la tabla que agrupa a esos 70 datos

--La **MEDIA** es la suma de todos los valores dividida por el número de ellos.
Símbolo: $\bar{x}$
Cálculo:
1) datos aislados, originales:

$$\bar{x} = \frac{\sum x}{N} \quad \text{; para el ejemplo A:} \quad \bar{x} = \frac{8+1+4+8+8+5+1}{7} = 5$$

$$\text{para el ejemplo B:} \quad \bar{x} = 39{,}6$$

2) datos agrupados en clases:

$$\bar{x} = \frac{\sum fc}{N} \quad \text{; en el ejemplo C:}$$

$$\bar{x} = \frac{(6*13)+(6*20)+(8*27)+(11*34)+(11*41)+(11*48)+(9*55)+8*62)}{70} = 39{,}4$$

Propiedades de la media

1- si a cada valor de x le sumamos, restamos, multiplicamos o dividimos por una constante, la media queda sumada, restada, multiplicada o dividida por esa constante
2- la media es sensible a la variación de cada valor de x
3- la media se expresa en la misma unidad de medida que los datos originales
4- si la media tiene decimales es habitual expresarla con uno más que los datos originales

Media aritmética ponderada

Se usa cuando se quiere o se debe dar una fuerza distinta a determinados valores.

$\overline{\mathbf{x}}_{\mathbf{pond}} = \frac{\sum \mathbf{xF}}{\sum \mathbf{F}}$ **, siendo x el valor original y F el factor de ponderación**

Ejemplos:

1) Al introducirse los estudios de Diplomatura en esta Escuela, el Área de Ciencias de la Enfermería englobaba diversas asignaturas, de cuyas notas salía la nota del Área. Como eran de extensión e importancia dispares, se decidió que Microbiología (que para abreviar llamaremos A) participaría con el 33%, la Bioestadística (B) con el 28%, las Prácticas (C) con un 23% y el resto, la media de Salud Pública, Organización e Historia de la Profesión ((D1+D2+D3)/3) conjuntamente con un 16%.
   Si las notas de las asignaturas fueron : 6 en A, 5 en B, 8 en C, 6 en D1, 8 en D2 y 10 en D3 , la nota del Área fué 6,5 y no la media aritmética 7,2
   $\overline{\mathbf{x}}_{\mathbf{pond}}$ = (6*33 + 5*28 + 8*23 + 8*16)/(33+28+23+16) = 6,5
2) la media de una distribución calculada a partir de una tabla es realmente una media ponderada en la que x es el punto medio de clase y f (frecuencia) el factor de ponderación F.

Otras medias

En circunstancias especiales (distribución con sesgo muy intenso) hay autores que prefieren otras medias como la media geométrica o la trimedia , en las que no vamos a entrar.
En los concursos varios jueces dan una nota al actuante. Para disminuir favoritismos e inquinas se utiliza la media recortada, que se obtiene prescindiendo del valor más alto y del más bajo. Este sistema se puede aplicar también para evitar errores, cuando se manejan grandes cantidades de datos y aparecen valores marginales "anómalos". Así se puede decidir no tener en cuenta un pequeño porcentaje (no más allá de un 3%) de los valores más altos y más bajos.

--La **MEDIANA** es el valor que ocupa el centro de la distribución una vez ordenados los datos. El símbolo es M

Cálculo:

1 – datos aislados, originales (¡que deben estar ordenados!)

a) N es impar: es el valor que ocupa el lugar (N+1)/2

b) N es par: es la media de los valores que ocupan los lugares N/2 y siguiente.

2 – datos agrupados

--de forma simplificada se toma como M el punto medio de la clase que contenga la mediana (el lugar se calcula como en los datos aislados) y se identifica la clase por la columna de frecuencias acumuladas.

--de forma un poco más exacta se utiliza la fórmula

$$M = L_i + i\left(\frac{N/2 - \sum f_M}{f_M}\right)$$

siendo L_i el límite inferior de la clase mediana, i su amplitud, N el nº total de datos, Σf_M las frecuencias acumuladas por debajo de la clase mediana y f_M la frecuencia de la clase mediana.

Ejemplos:

--supuesto A: se ordenan los 7 datos: 1 , 1 , 4 , 5 , 8 , 8 , 8 ; como N es impar la mediana será el valor que ocupe el lugar (7+1)/2 = 4 ; el 4º lugar es el 5

--supuesto B: se ordenan los 70 datos, número par. La mediana es la media de los valores que ocupen el lugar 70/2 = 35 y el siguiente, 36 . El 35º vale 40 y el 36º 41 , por tanto M = 40,5

--supuesto C: ***la clase mediana es la que contiene los valores 35º y 36º. En la columna de Σf se ve que pertenecen a la clase 38-44, que es la clase mediana. Por tanto M= c = 41

***aplicando la fórmula: $\mathbf{M = 37,5 + 7\left[\frac{\frac{70}{2} - 31}{11}\right] = 40}$

Propiedades de la mediana
Son las mismas que las de la media excepto la 2ª: la mediana sólo es sensible a la variación de los datos originales si se altera el orden en el centro de la distribución.

--La **MODA** es el valor más frecuente. Puede ocurrir que no haya moda o que haya más de una (empates en el máximo). El símbolo es Mo.
Cálculo:
-*en datos originales* se hace el recuento y se busca el valor más frecuente. Si hay empate, la moda es múltiple.
-*en datos agrupados* en tabla: la Mo será el punto medio de la clase modal, es decir, la más frecuente. En caso de empate se dan los puntos medios de las clases correspondientes.
Propiedades: como la mediana.
Ejemplos:
supuesto A: Mo = 8 ; supuesto B: Mo = 59 ; supuesto C: hay tres clases con frecuencia de 11; Mo = 34 , 41 y 48

De estos tres parámetros de tendencia central el mejor es sin duda alguna la media, pero hay algunos casos concretos (clases abiertas, valores muy discordantes) en que la mediana o incluso la moda son mejores. Cuando N≥30 la media suele ser un buen parámetro. En todo caso si el CV (coeficiente de variación), que luego veremos, supera el 50% la media no es buen representante del centro de la distribución.

Parámetros de dispersión

Informan de la dispersión de los datos, de la amplitud del conjunto. Los más importantes son:
-El RECORRIDO, que ya vimos en el tema 4 , o simplemente citar el máximo y el mínimo.
-La VARIANZA, que se basa en las diferencias entre cada valor y la media de la distribución.
-La DESVIACION ESTANDAR, que es la raíz cuadrada de la varianza.
-El COEFICIENTE DE VARIACIÓN, que relaciona la desviación estándar y la media.

--Varianza
Símbolo : s^2 (σ^2 , en la nomenclatura con caracteres griegos)
Cálculo: hay fórmulas distintas según los datos pertenezcan a una población o a una muestra.

– – población

– datos aislados: $s^2 = \frac{N \sum x^2 - (\sum x)^2}{N^2}$

– datos agrupados: $s^2 = \frac{N \sum (fc^2) - (\sum fc)^2}{N^2}$

– – muestra

– datos aislados: $s^2 = \frac{N \sum x^2 - (\sum x)^2}{N(N-1)}$

– datos agrupados: $s^2 = \frac{N \sum (fc^2) - (\sum fc)^2}{N(N-1)}$

Propiedades de la varianza

1- si a cada valor de x le sumamos o restamos una constante k, la varianza queda igual
2- si cada valor de x lo multiplicamos o dividimos por una constante k, la varianza queda multiplicada o dividida por k^2
3- la varianza es sensible a la variación de cada valor de x
4- la varianza se expresa en el cuadrado de la unidad de medida utilizada en la variable.
5- si la varianza tiene decimales, es habitual expresarla con dos decimales más que los datos originales

Ejemplos:
Con datos originales es conveniente construirse una tabla auxiliar con dos columnas: x y x^2.
--Así en el supuesto A (asumiendo que es una muestra):

x	x^2
8	64
1	1
4	16
8	64
8	64
5	25
1	1
--------------	--------------
35	235

$$s^2 = \frac{7 * 235 - 35^2}{7 * 6} = 10$$

--en el supuesto B : $s^2 = 207{,}58$
--en el supuesto C: la tabla auxiliar tendrá las columnas f , c , f*c , c^2 , fc^2 para que podamos tener los sumatorios necesarios para aplicar la fórmula.

$s^2 = 218{,}96$

--La **DESVIACION ESTANDAR** es la raíz cuadrada de la varianza y por tanto es un número más manejable y de utilización más frecuente.
Símbolo: s .También se usa mucho D.E. y la abreviatura inglesa S.D. Y la letra griega σ .

Fórmula: $s = \sqrt{s^2}$

Propiedades: como la media
Ejemplos:
-supuesto A: s = 3,2
-supuesto B : s = 14,4
-supuesto C: s = 14,8

--El **COEFICIENTE DE VARIACION** es un índice abstracto, que no tiene unidad de medida. Da igual que midamos la variable en cm , kg, sec., etc, , el coeficiente de variación se expresa siempre como %. (que puede ser mayor del 100%).
Símbolo: CV

Fórmula: $CV = \frac{100s}{\bar{X}}$

Aplicaciones:
1) comparar dispersiones de variables, incluso si están medidas en unidades distintas. La variable con el CV menor tiene la menor dispersión (y viceversa).
2) valorar la representatividad de una media. Es buena si no supera el 50%.

Ejemplos:
-supuesto A: 64%
-supuesto B: 36,4%
-supuesto C: 37,6%
-otro ejemplo: Los niños de 3 años de la ciudad C tienen una talla media de 93 cm con s = 3,8 . Los niños de 15 años de esa ciudad miden en media 162 cm con s = 6. ¿A que edad es la talla más variable?
Se calcula el CV: -a los 3 años: 4,09% -a los 15 años: 3,70%
Respuesta: La talla es más variable a los 3 años.

PARAMETROS DE FORMA

1) **SESGO** : es el grado de asimetría de una distribución, expresado por el coeficiente de sesgo o asimetría, cuyo valor ideal es 0 (entonces hay simetría). Cuando hay un Sesgo la parte más alta del histograma (o de la campana de Gauss) se desplaza hacia la derecha o la izquierda y la campana tiene una cola larga, donde estará la media, y otra más corta, en la que suelen estar la mediana y la moda. Si la media es menor que la M y/o la Mo, el sesgo es negativo y si es mayor, el sesgo es positivo.
Símbolo: Sg
Hay una fórmula, muy compleja, para calcular el coeficiente de sesgo, en la que no entramos.
Un cálculo aproximado es: $\mathbf{Sg} = \frac{\mathbf{3(\bar{x} - M)}}{\mathbf{s}}$, aunque lo mejor es observar la campana o el histograma. **Mirando la campana, si se desplaza a la derecha el sesgo es negativo; si lo hace a la izquierda, positivo**. Si nos ponemos en lugar de la campana, al revés.



Mirando el histograma de los “70 DATOS” (página 5.4) se ve que tiene un pequeño sesgo hacia la derecha, es decir, negativo. Con los datos originales el cálculo exacto da un sesgo de –0,196; la fórmula aproximada da -0,187. Con los parámetros calculados a partir de la tabla el sesgo vale según la fórmula aproximada –0,324.

2) **CURTOSIS**
es el grado de apuntamiento de una distribución, expresado por el coeficiente de curtosis, cuyo cálculo es complejo y no se ve aquí.
Símbolo: ct o k
Se toma como referencia a la campana de Gauss de la distribución normal, cuya k vale 0 y se dice que es mesocúrtica. Si la distribución es más alta y delgada, se dice que es leptocúrtica. y k es >0. Si es achatada y ancha se denomina platicúrtica y k es <0.
Los “70 DATOS” tienen una k = -1,105 y por tanto la distribución es algo platicúrtica.

PARAMETROS DE POSICION

1) **PERCENTILES**
Los percentiles (p) son parámetros de posición que nos indican la situación de cada valor en el conjunto de los datos ordenados, que se han dividido en 100 partes iguales. Se presentan como tabla o como gráfico.

Se expresan como pa siendo a el % de datos que queda por debajo del valor original al que corresponde ese percentil. Dicho de otra forma: a un valor le corresponde el percentil pa , cuando ordenados los datos el a% es menor que él y el (100-a)% es mayor.

Cálculo:

1- *en datos originales* : se ordenan los datos de menor a mayor y se calcula el lugar en el que estará el percentil (pa) buscado mediante la fórmula : lugar del pa = N*a/100. El valor que corresponda a es lugar o nº de orden será el pa
2- *en datos agrupados*: se utilizan la tabla o el gráfico de los porcentajes acumulados, interpolando, si es preciso. Hay una fórmula, parecida a la de la mediana, pero no suele ser necesaria.

Los percentiles se utilizan mucho en Pediatría en tablas y gráficos de crecimiento, pero en los últimos años su uso se ha extendido a muchos datos biológicos: colesterol, tensión arterial, densidad ósea... Han desplazado casi totalmente a otros parámetros de posición similares, como los deciles (el conjunto se divide en 10 partes iguales) y los cuartiles (el conjunto se divide en 4 partes).

Realmente hay100 percentiles, que van del p1 al p100, pero en la práctica se utilizan para mayor claridad sólo algunos de ellos. En Europa en las tablas y gráficos de crecimiento se utilizan el p3 , p10 , p25, p50, p75, p90, y p97.

El p50 se corresponde con el centro de la distribución: el 50% de los valores es mayor y el 50% es menor. Por tanto coincide con la mediana: p50 = M

En las variables biológicas los valores normales se obtienen a partir de muchas determinaciones en individuos sanos. Si un valor está por debajo del p3 se considera anormalmente bajo; si está por encima del p97, anormalmente alto; entre el p10 y el p90, totalmente normal. Entre el p3 y el p10, así como entre el p90 y el p97, aunque son aún normales, se consideran como en "zona de riesgo" o "sospecha", dada la proximidad de la zona anormal.

Los percentiles entre p25 y p75 pueden ser considerados también como parámetros de tendencia central y los mayores y menores como de dispersión.

Con los percentiles no pueden hacerse operaciones matemáticas, ya que son parámetros de posición . Así, pues, $p50 \neq (p25 + p75)/2$

Al final de este tema puede verse un ejemplo de gráficos percentilados del peso y talla de niños de 2 a 18 años. Un niño de 5 ½ años que pesa 23 kg y mide 106 cm tiene una talla en el p10, un peso <p90 y una relación peso/talla >p97.

2) La **PUNTUACION TIPIFICADA O NOTA TIPIFICADA** puede ser también considerada como un parámetro de posición. Se verá con detalle en el tema 10. Adelanto:

Símbolos: se utilizan varios según las escuelas: c, z, SDS , SDE...

Fórmula: $$\mathbf{c = \frac{X - \bar{X}}{s}}$$

Equivalencias aproximadas entre percentiles y puntuaciones tipificadas:

p	3	10	25	50	75	90	97
c	-2	-1,3	-0,7	0	0,7	1,3	2

Dos observaciones finales

1) una distribución queda perfectamente definida conociendo todos los parámetros que hemos visto. Como el sesgo y la curtosis son de cálculo más difícil, el mínimo son la media y la desviación estándar, que suelen anotarse así : $\bar{\mathbf{x}} \pm$ s ó $\bar{\mathbf{x}} \pm$DE .

Que la media sola no es suficiente lo aclara el clásico ejemplo del pollo:" si una persona se come dos pollos y otra no come ninguno, la Estadística dirá que se comen un pollo cada uno". La media es ciertamente 1 . Pero si calculamos la desviación estándar la valoración puede ser distinta:

-uno come 2 pollos y el otro ninguno:

x	x^2
2	4
0	0
---	---
2	4

$$\mathbf{s} = \sqrt{\frac{(2*4)-2^2}{2*1}}$$ = 1,4 y el CV = 140%

¡la media no es buena representante!

-cada uno come un pollo:

x	x^2
1	1
1	1
---	---
2	2

$$\mathbf{s} = \sqrt{\frac{(2*2)-2^2}{2*1}}$$ = 0 y el CV será 0%

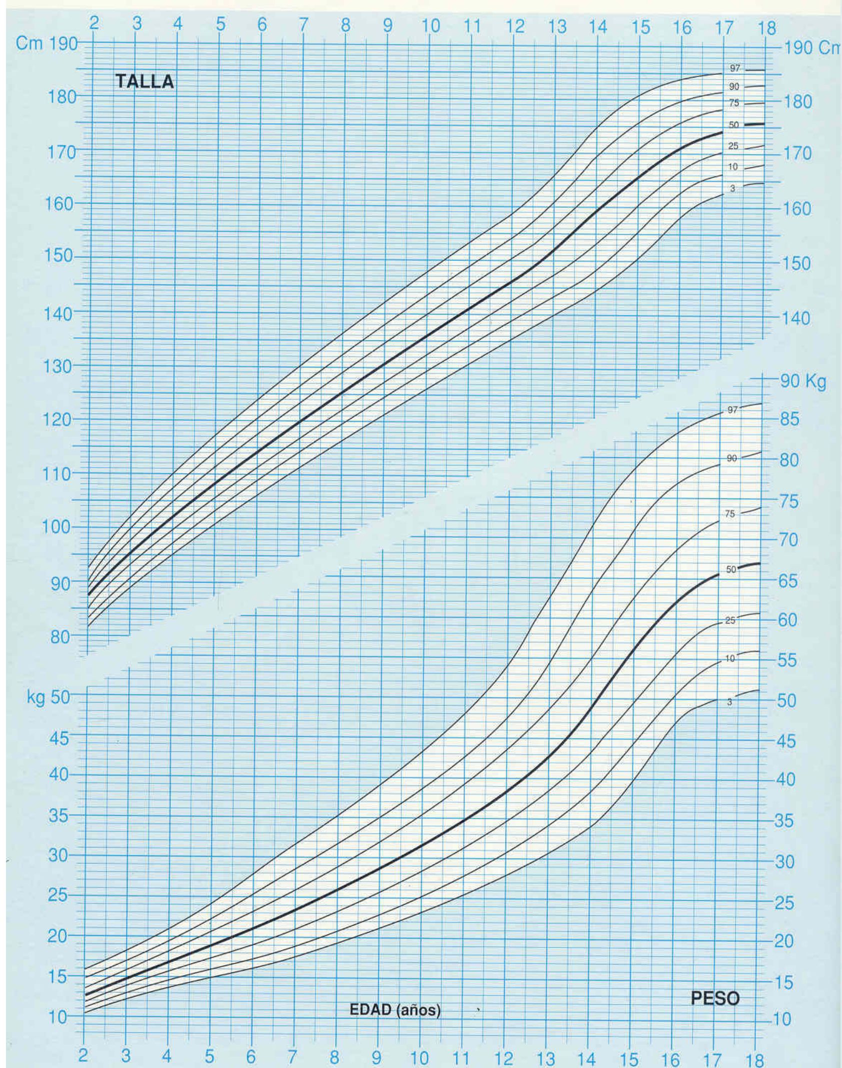
¡la media es buena representante!

2) siempre que sea posible, los índices se calcularán a partir de los datos originales, ya que los cálculos a partir de la tabla conllevan algo de error. Como puede verse en este resumen con parámetros de algunos ejemplos que se han ofrecido en este tema:

"70 DATOS"	**Datos originales**	**Datos agrupados**
Media	**39,6**	**39,4**
Desviación estándar	**14,4**	**14,8**
Mediana	**40,5**	**40**
Moda	**59**	**34 , 41 , 48**
Coeficiente de variación	**36,4%**	**37,6%**

NIÑOS: 2 a 18 años

TALLA
PESO

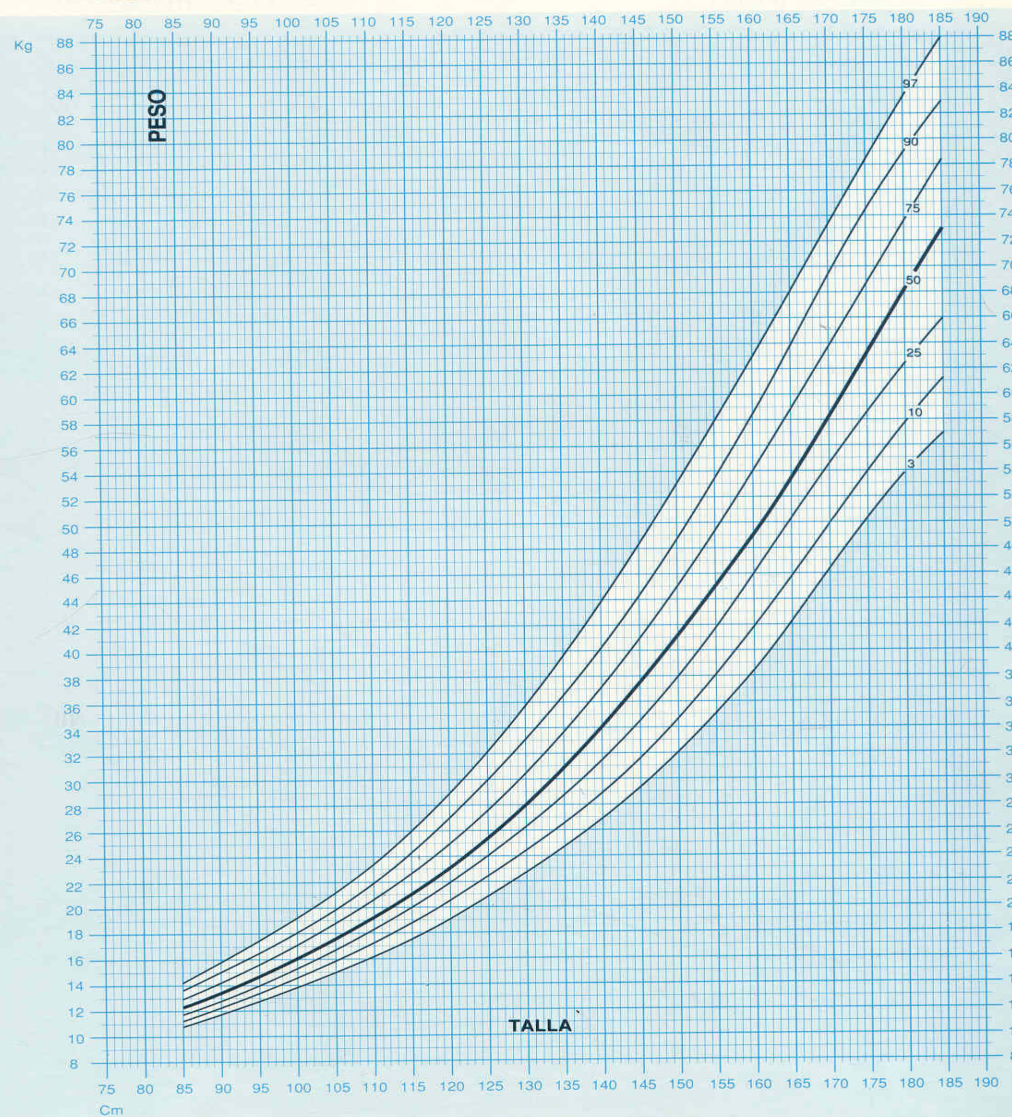


ESTUDIO LONGITUDINAL DE CRECIMIENTO. CURVAS DE 0 A 18 AÑOS.
M. Hernández, J. Castellet, J.L. Narvaiza, J. M. Rincón, I. Ruiz,
E. Sánchez, B. Sobradillo y A. Zurimendi.

INSTITUTO DE INVESTIGACION SOBRE CRECIMIENTO Y DESARROLLO.
FUNDACION F. ORBEGOZO. María Díaz de Haro, 10 bis, 48013 BILBAO
Realización Gráfica: Garsi Editorial - Londres, 17 - 28028 MADRID

NIÑOS

PESO-TALLA



ESTUDIO LONGITUDINAL DE CRECIMIENTO. CURVAS DE 0 A 18 AÑOS.
M. Hernández, J. Castellet, J. L. Narvaiza, J. M. Rincón, I. Ruiz,
E. Sánchez, B. Sobradillo y A. Zurimendi.

INSTITUTO DE INVESTIGACION SOBRE CRECIMIENTO Y DESARROLL
FUNDACION F. ORBEGOZO. María Díaz de Haro, 10 bis, 48013 BILBAO
Realización Gráfica: Garsi Editorial - Londres, 17 - 28028 MADRID

Tema 7 : DATOS BIVARIADOS. CORRELACION Y REGRESION.

Distribuciones uni- y pluridimensionales.

Hasta ahora se han estudiado los índices y representaciones de una sola variable por individuo. Son las distribuciones unidimensionales o univariadas .
En un individuo se pueden estudiar conjuntamente dos o más variables con objeto de ver si hay relación o dependencia entre ellas. Tenemos entonces distribuciones pluridimensionales, también llamadas plurivariadas. Cuando son dos se llaman bivariadas o bidimensionales. Son las únicas que veremos nosotros.
La simple medida de más de una variable en un individuo no tiene categoría de pluridimensional, sólo se tiene una serie de variables unidimensionales. ¡Hace faltar estudiarlas conjuntamente!

Estudio de variables bidimensionales

A una de las variables se la llama variable independiente y se representa por X. A la otra se la denomina variable dependiente y su símbolo es Y. (también se usan las minúsculas: x e y).
Los datos deben de ir siempre apareados. Para cada individuo se dan su X y su Y. ("Cada oveja con su pareja"). El nº de individuos se representa por N.

N es el nº de individuos, no el nº de datos, que siempre será el doble de N, pues cada individuo nos proporciona dos. ¡Es un error observado con frecuencia en los exámenes!

Ambas variables pueden ser cuantitativas (CT) o cualitativas (CL). En este tema veremos el caso de que ambas variables sean CT (que se completará en el tema 18) . En el tema 16 veremos la relación entre dos variables CL, expresada mediante la Odds ratio (OR). El caso de una variable CL y otra CT se trata en el tema 17.
--Ejemplos de variables bidimensionales
talla y peso, edad y tensión arterial, frecuencia cardiaca y frecuencia respiratoria, sexo y hábito de fumar, sexo y peso al nacer, velocidad de un vehículo y distancia de frenada...

Cuando ambas variables son CT, se pueden presentar:

a) cada variable por separado (con sus tablas, gráficos e índices)
b) conjuntamente (objeto de este tema) mediante:
    a. la tabulación y representación gráfica de los datos
    b. el cálculo de dos índices:
        i. coeficiente de correlación
        ii. ecuación de regresión

Tabulación

---de los datos originales

se hace una tabla, vertical u horizontal, con una columna (o fila) para X y otra para Y. Es opcional añadir otra para el número de orden del individuo. Los datos se ordenan en función del orden de los individuos o de los valores de X o de los valores de Y o no se ordenan en absoluto.
Ejemplo: Para X = (1 , 1 , 3 , 6 , 2 , 3 , 5 , 6) e Y = (1 , 1 , 4 , 4 , 2 , 5 , 1 , 5) :

Indiv.	X	Y
1	**1**	**1**
2	**1**	**1**
3	**3**	**4**
4	**6**	**4**
5	**2**	**2**
6	**3**	**5**
7	**5**	**1**
8	**6**	**5**

o

Ind.	1	2	3	4	5	6	7	8
X	**1**	**1**	**3**	**6**	**2**	**3**	**5**	**6**
Y	**1**	**1**	**4**	**4**	**2**	**5**	**1**	**5**

---de los datos agrupados en clases

Los valores de X e Y se agrupan en clases, siguiendo el método visto en el tema 4. La tabla es bidimensional: en la primera columna se representan las clases de X y en la primera fila las clases de Y. Al hacer el recuento los valores de cada individuo quedarán dentro de la casilla de la tabla que englobe a ambos.

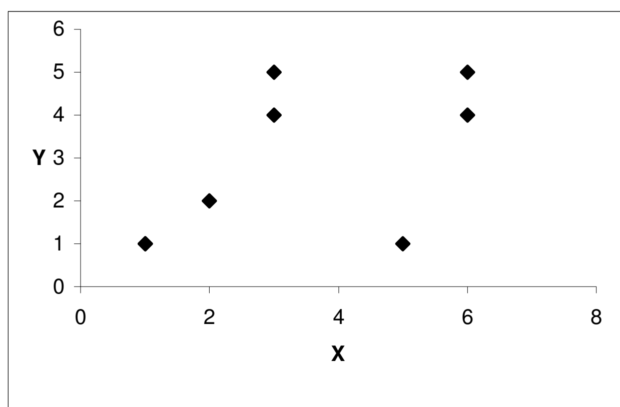
Ejemplo: Para los datos ya vistos la tabla podría ser así (presentada de forma simplificada y no del todo ortodoxa para mayor claridad):

X \ Y	1-2	3-4	5-6	TOTAL
1-2	3	0	0	3
3-4	0	1	1	2
5-6	1	1	1	3
TOTAL	4	2	2	8

Gráficos

--datos originales, aislados

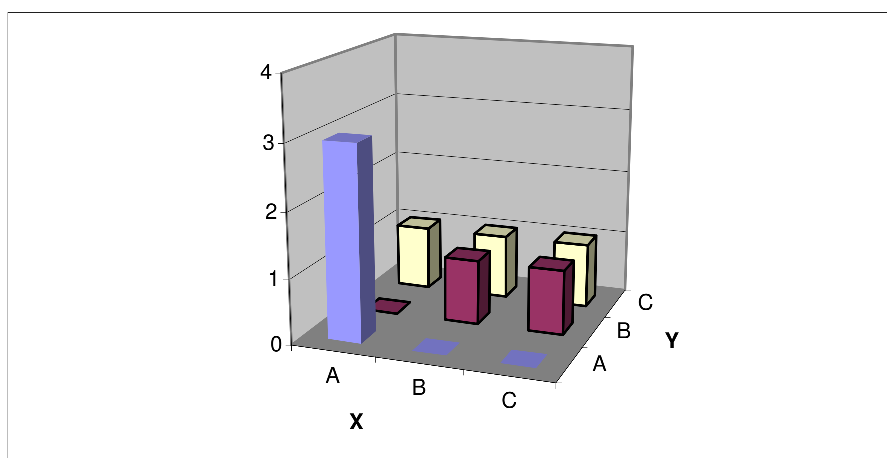
Es el diagrama de puntos, también llamado de dispersión o de nube de puntos. Los valores de cada individuo llevados aun eje de coordenadas originan un punto.



---datos agrupados en clases

El gráfico es el Estéreograma. Cada casilla de la tabla (que es la conjunción de dos clases, una de X y otra de Y) está representada por un prisma o cilindro (o incluso por una línea) cuya altura es proporcional a la frecuencia.

Para mayor claridad las clases en vez de como 1-2, 3-4 y 5-6 se representan como A, B y C



Índices estadísticos

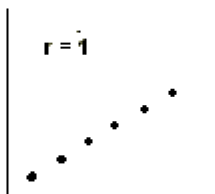
Los típicos de estas distribuciones, aparte de los de cada variable por separado, son el coeficiente de correlación y la ecuación de regresión. Son los llamados índices o parámetros de asociación. Son distintos en función del tipo de variables (CL-CL, CL-CT, CT-CT). en este tema sólo nos ocuparemos del caso en que ambas variables son CT.
Correlación significa relación mutua y expresa el grado de asociación existente entre las variables, el CUANTO de la relación. Su parámetro es el coeficiente de correlación. Su símbolo es **r**, que puede acompañarse, si la claridad lo exige, de un subíndice con la notación de las variables (p.e. r_{xy}). Se puede calcular la correlación entre dos variables o más (correlación múltiple).
La **regresión** es la forma, el COMO de esa asociación. Expresa la relación entre las dos variables, X e Y, mediante la ecuación de regresión y su representación gráfica la línea de regresión. Mediante ella conocida una variable es posible predecir la otra. Por consenso X es la variable independiente e Y la dependiente. De esta forma $Y = f(X)$.

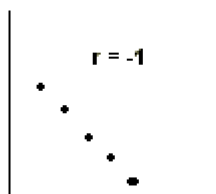
Coeficiente de correlación

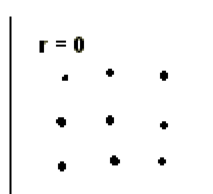
Mide la intensidad de la asociación entre las variables. Es un número abstracto, independiente de la unidad de medida de las variables. Puede adoptar cualquier valor entre –1 y 1. Dicho de otra forma: $r = \in (-1 \div 1)$. Suele expresarse con 3 decimales, a no ser que valga –1, 0 ó 1. Aparte de su valor descriptivo sirve para ver la significación estadística de la relación (tema 18)
Aquí veremos sólo la correlación entre dos variables. Su coeficiente de correlación se llama de Pearson, aunque cuando se dice simplemente coeficiente de correlación, se sobreentiende que es éste. En el tema 18 se verá otro coeficiente, el de Spearman, que se usa cuando no puede utilizarse el de Pearson.
Si se observa una correlación aparentemente alta entre X e Y puede tratarse de dos situaciones:
--una variación de X provoca otra en Y. Por ejemplo, el aumento de la temperatura corporal produce un aumento de la frecuencia cardiaca.
--X e Y varían a la par por efecto de un a tercera o más variables. La correlación existente es pura coincidencia. Son las llamadas correlaciones espurias, ya citadas en el tema 1. Son las más frecuentes. De forma automática correlación ≠ causalidad. Se requiere un estudio experimental con resultado significativo.

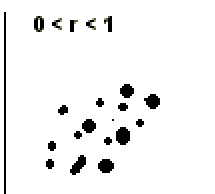
Si $r = 1$ hay una correlación total (perfecta) positiva.
Si $r = -1$ hay una correlación total (perfecta) negativa.
Si $r = 0$ no hay correlación.
Si está entre –1 y 0 , la correlación es parcial y negativa.
Si está entre 0 y 1, la correlación es parcial y positiva.
Una r de 0, -1 ó 1 apenas se encuentra en la práctica

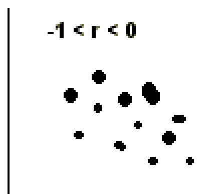
Gráficamente esto se puede representar así:











Cálculo de coeficiente de correlación
Veremos únicamente el cálculo a partir de los datos originales, aislados.

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

Para este cálculo y el de la ecuación de regresión es de gran ayuda construirse una tabla auxiliar como la que se utiliza en el siguiente ejemplo:

X = (2 , 1 , 3 , 2 , 5) ; Y = (3 , 5 , 4 , 2 , 6)

X	**Y**	**X^2**	**Y^2**	**XY**
2	3	4	9	6
1	5	1	25	5
3	4	9	16	12
2	2	4	4	4
5	6	25	36	30
13	*20*	*43*	*90*	*57*

$$\mathbf{r} = \frac{(5 * 57) - (13 * 20)}{\sqrt{[(5 * 43) - 13^2][(5 * 90) - 20^2]}} =$$

$$= \frac{25}{\sqrt{46 * 50}} = 0,521$$

Este valor de r es el valor puntual. Cada día se utiliza más el valor por intervalo, cuyo cálculo veremos en el tema 13, en el que se estudian los intervalos de confianza (IC).

Regresión

Ya hemos visto el concepto de regresión. La fórmula matemática que la expresa puede ser una ecuación de primer grado (regresión lineal: y = a+bx) u otras ecuaciones más complejas (cuadrática: $y=ax^2+bx+c$; exponencial: $y=ae^{bx}$; potencial: $y=ax^b$; hiperbólica: y=a(b/x) ; logarítmica: $y=a+bl_nx$; etc...), que no trataremos, pues son muy complejas. Nos limitaremos a la regresión lineal, también llamada recta de regresión, pues su representación gráfica es una línea recta, que representa lo mejor posible a todos los puntos del diagrama de dispersión. Realmente se podrían trazar muchas rectas de regresión, pero sólo nos interesa la llamada "mejor línea de ajuste", que es la que corresponde a la ecuación y=a+bx (ó y=bx+a; el orden de los sumandos no altera la suma).

En esta fórmula **b** es el coeficiente de regresión, también llamado pendiente, pues de él depende la inclinación de la recta y nos indica en cuanto se modifica **y** en media cuando **x** varía en una unidad.

a es el valor de y cuando x = 0 , por lo que también se la llama ordenada en el origen o intersección de y . Se ha comprobado que la mejor línea de ajuste es aquella en que la suma de los cuadrados de las diferencias entre cada punto original y la línea de regresión es la menor de todas las posibles. Por eso a este método se le llama "de los mínimos cuadrados". Afortunadamente no hay que calcularlos, pues se ha desarrollado una fórmula mucho más manejable para encontrar la ecuación.
En principio se considera a y variable dependiente y a x variable independiente, por lo que la regresión se dice que es de y sobre x. En este sentido b es realmente b_{yx} y así se entiende cuando no hay subíndice. Matemáticamente también se puede calcular la regresión de x sobre y. Si interesara este cálculo, lo que no es habitual, escribiríamos b_{xy} para evitar confusiones.

Cálculo
Seguiremos el procedimiento que calcula primero **b** y a partir de él calcula **a**

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2} \qquad a = \overline{Y} - b\overline{X}$$

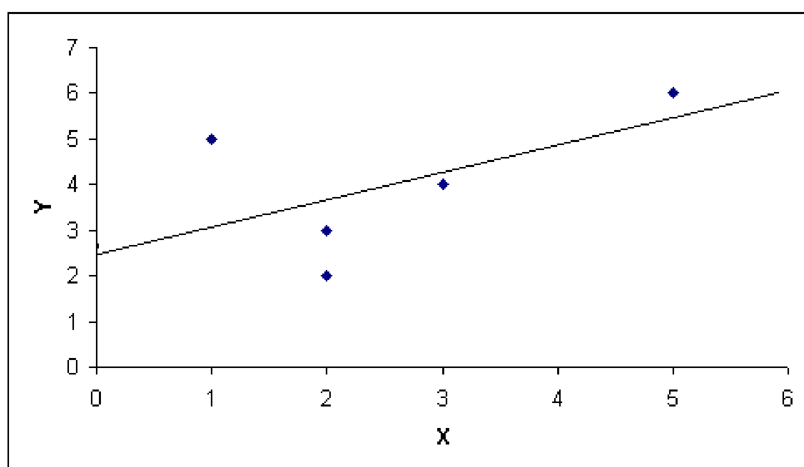
Ejemplo: Utilizando los datos empleados para calcular el coeficiente de correlación:

$$\overline{X} = \frac{13}{5} = 2,6 \qquad \overline{Y} = \frac{20}{5} = 4 \qquad b = \frac{(5*57)-(13*20)}{(5*43)-13^2} = \frac{25}{46} = 0,54347$$

$$a = 4-(0,54347*2,6) = 2,587$$

por tanto la ecuación es y = 2,587 + 0,543x

Representación gráfica
Para trazar una recta basta con dos puntos. En el diagrama de dispersión se busca el valor de y para x = 0. El otro punto se obtiene a partir de un valor cualquiera de **x** que nos de una **y** que no se salga del gráfico. En nuestro ejemplo: si x = 0 , y = 2,587 ; para x = 5 , y = 5,302

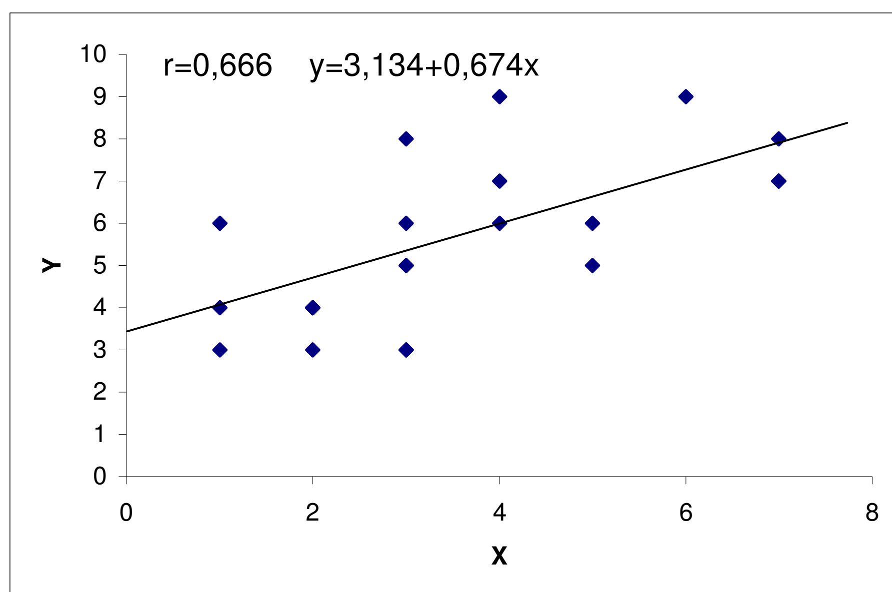


Se suele incluir en el gráfico la ecuación y el coeficiente de correlación y con menos frecuencia el IC (intervalo de confianza) de forma numérica y/o con dos rectas más que lo delimiten.

Coeficiente de determinación
Mide cuantitativamente la bondad o representatividad del ajuste de la recta a la nube de puntos. Es el cuadrado de r. Su símbolo es $\mathbf{r^2}$ o R. En nuestro ejemplo r^2 = 0,302 . Cuando se calculan diversas ecuaciones de regresión (lineal, exponencial, logarítmica, etc.) la que tenga el r^2 más alto será la mejor, la más representativa. $\mathbf{r^2}$ unifica la fuerza de la asociación de positivos y negativos. (una r = -0,400 es más potente que una r = 0,350 ; sus $\mathbf{r^2}$ son 0,160 y 0,122)

Ejercicio resuelto con Excel.
Ejercítese en el cálculo de la media, desviación estándar, CV, coeficiente de correlación y ecuación de regresión.

X	Y				X	Y
2	4	**Error est. Y**	1,472	**media**	3,478	5,478
3	3	**r**	0,666	**s**	1,904	1,928
5	5	**Ecuación: b**	0,674	**CV**	54,7	35,2
2	4	**Ecuación: a**	3,134	**p50 ó M**	3,000	5,000
1	3					
2	4	**N = 23**				
7	7					
3	6					
2	3					
4	6					
1	6					
3	3					
1	4					
3	5					
2	4					
6	9					
4	9					
3	8					
4	7					
3	5					
5	6					
7	7					
7	8					



Notas adicionales. 1) Con los datos del ejercicio anterior se han calculado otras ecuaciones de regresión con sus respectivos r y r^2. Se dan aquí a título puramente informativo para que se vea que la mejor ecuación que relaciona a X e Y es la cuadrática, ya que tiene la r^2 más alta.

ECUACION	a	b	c	r	r^2
Cuadrática	**-0,034**	**0,950**	**2,703**	**0,668**	**0,447**
Lineal	**3,134**	**0,674**		**0,666**	**0,443**
Exponencial	**3,334**	**0,125**		**0,659**	**0,434**
Logarítmica	**3,262**	**2,034**		**0,630**	**0,397**
Potencial	**3,412**	**0,378**		**0,625**	**0,390**

2) aunque no es lo correcto, en la práctica se calcula en ocasiones **r** cuando se contrastan 2 Vbles. CT procedentes de individuos distintos, siempre que estén emparejados. Aquí N es el nº de parejas de datos, no el de individuos.

Tema 8 : Series de tiempo

Concepto

Una serie de tiempo representa las variaciones o evolución de un fenómeno a través del tiempo. Se concreta en una serie de observaciones de una variable, hechas en determinados intervalos de tiempo, generalmente iguales. Son datos bivariados en los que la variable independiente es el tiempo, que se simboliza por t en vez de por x.
Son muy utilizadas en la vida diaria: evolución en un determinado periodo de tiempo de la producción de coches, exportaciones, turistas que nos visitan, paro, etc. La clásica curva de la fiebre y pulso de un paciente es una serie de tiempo. Los modernos monitores de las llamadas constantes vitales y los barógrafos, termógrafos y aparatos similares hacen un registro continuo de una o más variables.

Representación

a) de forma numérica o tabular. La columna base es el tiempo.
b) de forma gráfica. La más usada es el diagrama lineal, la variante del polígono de frecuencias que no baja al eje de abscisas ya que no se abarca toda la distribución sino sólo una parte de la misma. Si abarca toda la distribución se usará el polígono de frecuencias. En el eje de abscisas se representa el tiempo y en el de ordenadas la frecuencia correspondiente.

La tabla puede acompañarse de una columna con números índice, que en general parten de considerar como 100 ó 100% al valor de Y en el primer periodo de tiempo. Para los demás periodos se hace el cálculo por una simple regla de tres. También puede ponerse una columna que represente una tasa.
Ejemplo:

HOSPITAL H			
Ingresos del Servicio S			
año	ingresos	Nº índice	tasa/100.000 hab.
2000	800	100	200
2001	915	114	229
2002	980	122	245
2003	1040	130	260
2004	1000	125	250
2005	980	122	240

Otros cálculos

Los más utilizados son el coeficiente de correlación y la ecuación de regresión.

Lo esencial de las series de tiempo

Su estudio ha permitido comprobar que están sometidas a **variaciones típicas**, siendo las más importantes las tres siguientes:
--variaciones a largo plazo o tendencia secular. Representan la variación general de la serie, suavizada por la absorción de otras variaciones menores en intervalos de tiempo largos. Podría decirse que los datos utilizados son medias de otros muchos datos. Un ejemplo típico es la talla media de los chicos españoles cuando se incorporaban al servicio militar obligatorio, registrada durante casi un siglo.
--variaciones a medio plazo o fluctuaciones periódicas, obtenidas en intervalos de tiempo menores. Pueden ser estacionales y cíclicas. Son estacionales cuando el plazo es menor de un año. Ejemplo típico son las ventas de unos grandes almacenes en Navidad-Reyes, San Valentín, Día de la Madre, etc. Las cíclicas ocurren a intervalos mayores de un año, como los ciclos de la economía. Suelen ser más suaves.

--variaciones irregulares o accidentales. No son previsibles, como el aumento de las ventas de determinados alimentos cuando se rumorea que van a subir mucho de precio o la disminución de la producción de una fábrica durante una huelga. Estas variaciones pueden originar nuevos ciclos o tendencias, como la crisis pesquera de los años 70, que elevó mucho los precios, sin vuelta atrás. O el aumento imparable del precio del petróleo tras la primera invasión de Irak.

Análisis de las series de tiempo

Es una especialidad de la Estadística. No podemos entrar en sus procedimientos, pues son muy complejos y desbordan las posibilidades de tiempo de esta asignatura. Únicamente veremos sus aplicaciones. Las principales son:
--descripción y estudio de un fenómeno a lo largo del tiempo con todas sus variaciones.
--predicción de la tendencia para el futuro. Se basa en la ecuación de regresión, mejor con su intervalo de confianza, lo que da una horquilla de posibles situaciones. Aquí hace falta una buena dosis de experiencia y sentido común. Utilizando la ecuación de regresión de la mortalidad de una enfermedad en los primeros años tras introducir una vacuna eficaz, se puede llegar fácilmente a una mortalidad negativa, es decir, a la resurrección de los muertos...

Precauciones

Las series de tiempo se prestan mucho a la manipulación. Por ejemplo utilizando variaciones cíclicas, o incluso accidentales, como si fueran tendencias a más largo plazo. O tomando como punto de partida de la serie un "momento conveniente" para lo que interesa. Valorarlas siempre con espíritu crítico.

Otro ejemplo:

Hotel del Golfo

Estancias agosto últimos 5 años

Año	Estancias	Nº índice
2001	2980	100.0
2002	3050	102.3
2003	3130	105.0
2004	3020	101.3
2005	3260	109.4

r = 0,757

Y = 48,2 *X – 93420,2

o sea, **Estancias = 48,2*año – 93420,2**

Predicciones:
año 2006: Estancias = 48,2*2006 - 93420,2 = 3269
año 2007: Estancias = 48,2*2007 - 93420,2 = 3317

Tema 9 : Teoría de la probabilidad

Definición

Veremos dos:

---La **definición clásica de Laplace** dice que la probabilidad, (p), de ocurrencia de un fenómeno A (o evento, suceso, modalidad de una variable...) en un experimento aleatorio de resultados equiprobables es igual al nº de casos favorables, también llamados éxitos, (símbolo: f ó r) dividido por el nº de casos posibles (N).

$$pA = f/N$$

Como f puede estar entre 0 y N, los valores posibles de p van de 0 a 1. Suelen expresarse, salvo el 0 y el 1, con 3 ó 4 decimales. También se puede expresar como porcentaje, entre 0% y 100%. A veces es conveniente, por ser más manejable, expresarlo como fracción.

Tres aclaraciones a esta definición

1-Un *experimento aleatorio*

- -no tiene resultado fijo, sino un conjunto de posibles resultados (2 ó más)
- -el resultado no se conoce de antemano, ocurre de forma aparentemente casual.
- -se puede repetir indefinidamente bajo las mismas condiciones.

2- *Equiprobable* quiere decir que todos los resultados tienen la misma probabilidad de ocurrir

Ejemplo: la probabilidad de que al tirar un lado salga un 3 es 1/6 .(1/6 es preferible a 0,1667).
El modelo de Laplace es un modelo teórico, intuitivo, en el que por simple reflexión se pueden saber las probabilidades.

3- *Éxito* se utiliza cuando ocurre el evento. El término es un clásico y se introdujo estudiando tiradas de dados, aplicándose aunque el evento sea algo negativo. Si se estudia la mortalidad, un fallecimiento será un "éxito"...

---La **definición de Richard von Misses** es más amplia y universal, basada en un modelo experimental, práctico: "La mejor estimación de la probabilidad de la ocurrencia de un fenómeno en un experimento aleatorio es su frecuencia relativa".

Ejemplo:. Teóricamente al lanzar una moneda bien hecha la p de cara es de 0,5. Hacemos un experimento tirando la moneda repetidamente. Vamos anotando como éxito las caras que van saliendo y después de cada tirada se calcula la f.r. de éxitos. Tras variaciones de cierta amplitud al principio pronto la f.r. se mueve cada vez más cerca de 0,5, con el que coincidirá exactamente en el infinito.

De esta forma calculando la f.r. podemos hallar la probabilidad de sucesos en los que no podemos utilizar la intuición. Por ejemplo, tirando varios cientos de chinchetas del modelo X al suelo, la f.r. de las que queden con la punta hacia arriba nos dará la p de tal resultado en ese modelo.

No tiene valor estadístico la llamada probabilidad subjetiva, que es una mezcla del conocimiento de los factores que pueden influir en un resultado con factores emocionales. Como la p de que nuestro equipo favorito gane el próximo partido o de aprobar una asignatura a la primera..

Sucesos elementales y complejos

Suceso elemental es el suceso básico, como p .e. nacer chica, cuya p es de 0,5

El suceso complejo comprende varios elementales, como p.e. tirar dos dados o el nº de chicas en una familia de 5 hijos. En algunos casos es fácil calcular sus probabilidades de ocurrencia con las reglas que se ven a continuación, pero en la mayoría hay que recurrir a las distribuciones fundamentales de probabilidad, que se verán en el tema 10

Algunos conceptos básicos de la probabilidad

1- $0 \leq p \leq 1$ ó $0\% \leq p \leq 100\%$

2- $\Sigma p(A_x) = 1$, siendo A_x el dominio de la variable, o sea todas sus modalidades o valores

3- Si A es el suceso elemental con probabilidad pA, la probabilidad de que no ocurra A, es decir, de que ocurra el suceso contrario o complementario ($\bar{A}$) es 1-p ó q .

Por tanto $p\bar{A} = 1-p = q$; $qA = 1 - q$

Un suceso elemental y su complementario son mutuamente excluyentes, incompatibles, no pueden ocurrir simultáneamente. Un suceso complementario puede ser simple o múltiple. Simple o sencillo, cuando sólo tiene una modalidad (caso de una moneda). Múltiple o compuesto, cuando engloba varias modalidades (caso de un dado).

4- $p + q = 1$ ó $p + q = 100\%$

5- Son sucesos independientes aquellos cuya ocurrencia no depende de otro u otros sucesos. Por ejemplo, que al tirar dos dados en una salga 4 y en el otro 2.

Son sucesos dependientes aquellos cuya ocurrencia depende de otro u otros sucesos. Si sacamos dos cartas de una baraja española, la p de que la segunda sea oros depende del palo de la primera carta. se formula así: p (A2/A1) , "p de A2 dado A1".

6- Ley multiplicativa. Rige la p de que ocurran a la vez dos o más sucesos (que por fuerza tienen que ser compatibles).

a. si son independientes: p(A1 y A2) = pA1 * pA2
b. si son dependientes: p(A1 y A2) = pA1 * p(A2/A1)

7- Ley aditiva. Rige la p de que ocurra un suceso u otro.

a. si son incompatibles. p(A1 o A2) = pA1 + pA2
b. si son compatibles: p(A1 o A2) = pA1 + pA2 – pA1*pA2
   ya que hay que restar la compatibilidad.

<u>Ejemplos</u>

a) p de que al tirar un dado dos veces salgan en ambas un 6.
   "seis en la 1ª tirada **y** 6 en la 2ª"
   p(2 veces 6) = 1/6 * 1/6 = 1/36 (mejor que 0,0278)
b) p de que al tirar dos dados salga en ambos un 6
   "seis en el primer dado **y** seis en el segundo"
   es el mismo caso que a)
c) La p de ser rubio es de 0,3 y la de llevar gafas es de 0,2 . Calcular la p de que una persona cualquiera sea rubia y lleve gafas (se asume que son independientes)
   p(rubio **y** gafas) = 0,3 * 0,2 = 0,06 (ó 6%)
d) en una caja hay 3 bolas blancas y 2 negras. Calcular la p de que sacando dos bolas, las dos sean negras.
   Nos piden la p de que sea negra la primera **y** negra la segunda.
   la p de ser negra de la 1ª bola es 2/5 ; una vez sacada quedan 4 bolas (una, negra)
   la p de ser negra de la 2ª bola es de ¼
   p(2 bolas negras) = 2/5 * ¼ = 2/20 = 1/10 (ó 0,1 ó 10%)
e) p de que al sacar una carta de una baraja española de 40 cartas sea oros o copas.
   p(oros **o** copas) = 10/40 + 10/40 = 20/40 = ½ (ó 0,5 ó 50%)
f) p de que al sacar una carta de esa baraja sea as o espadas.
   hay 4 ases , 10 espadas y 1 as de espadas (que cuenta como as y como espada, 1 entre 40, que debe ser compensada)
   p(As **o** Espada) = 4/40 + 10/40 – 1/40 = 13/40 = 0,325
g) p de acertar 6 en la Primitiva
   Hay 49 bolas. Como no hay reemplazo, cada vez que sale una bola, queda una menos en el bombo. Para acertar los 6 resultados hay que acertar el primer número y el segundo y el tercero...y el sexto.
   p(6 aciertos) = 6/49 * 5/48 * 4/47 * 3/46 * 2/45 *1/44 = 1 /13.983.816
h) p de que tirando un dado 4 veces, la primera vez que salga un 5 sea en la 4ª tirada.
   p(5 sólo en la 4ª) = p(no 5 en la 1ª)*p(no 5 en la 2ª)*p(no 5 en la 3ª)*p(5 en la4ª)
   = 5/6 * 5/6 * 5/6 * 1/6 = 125/1296 = 0,096

a) p de al menos un éxito (es decir, uno o más, uno como mínimo) en n intentos

se resuelve así : $p(r\geq 1) = 1 - p(r=0)^n$ ¡Ojo! no es $1 - p(r=0)*n$

Ejemplo: Un problema importante en la prevención del tétanos cuando no había vacunas o gammaglobulinas y había que administrar suero antitetánico eran las reacciones, a veces muy graves, que ocurrían en un 10% de los inyectados.

En una persona que hubiera recibido 10 inyecciones ¿cual es la p de que al menos tuviera una reacción?

Si la p de tener una reacción es de 0,1, la de no tenerla es de 0,9. Por tanto $p(r\geq 1) = 1 - 0{,}9^{10} = 0{,}651$. Si, falsamente, se hubiera calculado 1 – 0,9*10 se obtendría un resultado imposible: p = -8

Distribución de probabilidad

es el conjunto de las p de todas los valores o modalidades que puede adoptar una variable X. Veamos el caso más sencillo, el de una variable cualitativa:

--se establece el dominio de la variable (todas las modalidades)

--se calcula la p de cada modalidad

--se tabula y se representa gráficamente

<u>ejemplo</u>: X = suma de puntos al tirar dos dados

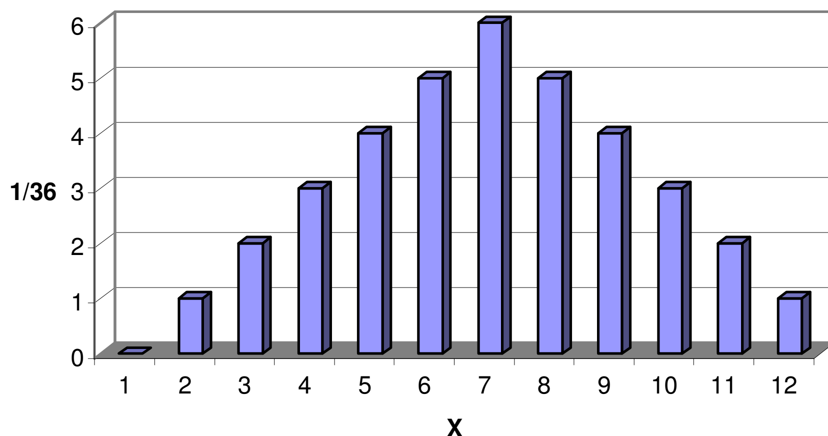
dominio: hay 36 combinaciones posibles (zona sombreada)

		dado 1					
		1	*2*	*3*	*4*	*5*	*6*
d	*1*	**2**	**3**	**4**	**5**	**6**	**7**
a	*2*	**3**	**4**	**5**	**6**	**7**	**8**
d	*3*	**4**	**5**	**6**	**7**	**8**	**9**
o	*4*	**5**	**6**	**7**	**8**	**9**	**10**
	5	**6**	**7**	**8**	**9**	**10**	**11**
2	*6*	**7**	**8**	**9**	**10**	**11**	**12**

probabilidad:

x	1	2	3	4	5	6	7	8	9	10	11	12
px	0	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

gráfico :



Método de Bayes

El modelo estadístico bayesiano se basa en probabilidades condicionadas y ha permitido el desarrollo, aún bastante imperfecto, del "diagnóstico por ordenador". A partir de las frecuencias de determinados síntomas en diversas enfermedades calcula la p de padecer una u otra enfermedad. Es un compleja especialidad dentro de la Estadística, cuyos detalles escapan a la intención de esta asignatura. Veremos su fórmula general y un ejemplo.

Fórmula de Bayes

$$p(A_x / E) = \frac{pA_x * p(E / A_x)}{\sum_{i=1}^{n} \left[pA_i * p(E / A_i) \right]}$$

pudiendo valer x entre 1 y n

Ejemplo

Se sabe que la presencia de determinados síntomas se da en el 60% de pacientes con la enfermedad A1, en el 30% de los que padecen la enfermedad A2 y en el 10% de los que tienen la enfermedad A3.
Al análisis E sale positivo en el 30% de los casos de A1, en el 70% de los casos de A2 y en el 70% de los de A3.
Si un paciente tiene esos síntomas y el análisis sale positivo, ¿qué probabilidades hay de que tenga una u otra enfermedad?

Enferm. (Ax)	pAx	p(E+/Ai)	pAx*p(E/Ai)	pAx/ E+
A1	0,6	0,3	0,18	0,18/0,46 = 0,391 = 39,1%
A2	0,3	0,7	0,21	0,21/0,46 = 0,456 = **45,6%**
A3	0,1	0,7	0,07	0,07/0,46 = 0,152 = 15,2%
		Suma .	0,46	

La enfermedad más probable es la A2, seguida de cerca por la A1 y más lejos por la A3.

Tema 10 : Distribuciones fundamentales de probabilidad

Ya hemos visto que los fenómenos naturales siguen el modelo indeterminista, es decir las leyes del azar, entendido como la combinación de múltiples factores, en gran parte desconocidos e incontrolables, que conducen a resultados no previsibles de antemano, aunque sí conocidos, que se caracterizan por su variabilidad en los diferentes individuos. A cada uno de los posibles resultados se asocia una probabilidad, que en sucesos sencillos o poco complejos es fácil de calcular por las leyes básicas o fundamentales de la probabilidad, pero al aumentar la complejidad el cálculo se hace muy difícil o imposible. Entonces hay que recurrir a una serie de modelos teóricos, las llamadas distribuciones o leyes fundamentales de la probabilidad, que nos permiten hacer el cálculo con relativa facilidad. Al aumentar el nº de individuos todas las distribuciones se van aproximando y acaban confluyendo y haciéndose una en el infinito.

Clasificación

a) para variables discretas
   - --D. binomial
   - --D. polinomial
   - --D. de Poisson
   - --D. hipergeométrica

b) para variables continuas
   - --D. normal
   - --D. de la t de Student
   - --D. de la χ^2 de Pearson
   - --D. de la F de Snedecor-Fisher

Para todas valen los principios que ya conocemos:

$0 \leq p \leq 1$

$p + q = 1$

$\Sigma p(x) = 1$

En este tema nos ocuparemos de las distribuciones binomial, de Poisson, normal y hipergeométrica . En el Anexo se verán la t de Student, la χ^2 y la F. No veremos la polinomial.

DISTRIBUCION BINOMIAL

Concepto

es el modelo básico de distribución de las variables discretas (o discretizadas), que como ya sabemos pueden ser reducidas en última instancia a dicotómicas.

Experimentos binomiales

Pueden ser elementales y complejos

Los **elementales** tienen dos resultados posibles: Éxito (cuando aparece el resultado que se pretende) y fracaso , que puede ser único o múltiple. Sus probabilidades respectivas son p y q

En los **complejos**
- --el experimento elemental se repite n veces
- --obteniendo r éxitos (de 0 a n) : $0 \leq r \leq n$
- --cada modalidad de la variable va asociada a una r . Como r empieza en 0 siempre hay n+1 modalidades: la de r=0 y las de r entre uno y n.
- -- un experimento binomial complejo puede repetirse **N** veces. Cada modalidad aparecerá **Nr** veces.

Notación

La distribución suele designarse como DB, pero cuando se dan los parámetros típicos, la n y la p del suceso elemental, se utiliza sólo B . Así: **B(n , p)**

Algunos ejemplos:

Experimento	Éxito	p	n	r	notación
elemental: lanzar 1 moneda complejo: lanzar 4 monedas	salir cara salir cara	0,5 0,5	1 4	0 , 1 0 , 1 , 2 , 3 , 4	B(1 , 0,5) B(4 , 0,5)
elemental: lanzar un dado complejo: lanzar 5 dados	salir 1 salir 1	1/6 1/6	1 5	0 , 1 0 , 1 , 2 , 3 , 4, 5	B(1 , 1/6) B(5 , 1/6)
elemental: familia con 1 hijo complejo: familia con 4 hijos	ser chica ser chica	0,5 0,5	1 1	0 , 1 0 , 1 , 2 , 3 , 4	B(1 , 0,5) B(4 , 0,5)

El lanzamiento de las 4 monedas se puede repetir N veces.
O podemos estudiar N familias de 5 hijos.

Cálculo de las p de r

1) fórmula $$p(r) = \binom{n}{r} p^r q^{n-r} = \frac{n!}{r!*(n-r)!} p^r q^{n-r}$$

$\binom{n}{r}$ da los coeficientes del desarrollo del binomio de Newton

2) tablas (en la pagina 16 hay una para $n \leq 8$ y ciertos valores de p)

3) Método intuitivo (la clásica “cuenta de la vieja”) posible en algunos casos.

Gráfico : diagrama de barras

Otros parámetros

Media o esperanza matemática: $\bar{X} = np$
la media representa el nº esperado de éxitos en el experimento
Varianza: $s^2 = npq$

y por tanto, **desviación estándar**: $s = \sqrt{npq}$

<u>**n , p , N y Nr**</u>
conviene insistir en estos símbolos que son básicos en la DB.
n : veces que se repite el suceso elemental en un experimento binomial. Si n=1 es un experimento simple; si >1, es complejo
p : probabilidad del suceso elemental
N : veces que se repite el experimento complejo. Si no se dice nada, N=1
$\mathbf{N_r}$: frecuencia de cada modalidad tras N repeticiones. $\mathbf{\Sigma N_r = N}$
----Si <u>tiramos una moneda 1 vez</u>, es una B(1 , 0,5) . Podemos obtener 0 ó 1 cara (r). N=1
Si <u>este experimento lo repetimos 3000 veces</u> (N) seguirá siendo una B(1, 0,5) pero con N=3000. r sigue valiendo 0 y 1. Nos pueden salir p.e. 1450 caras. Entonces $N_0 = 1550$ y $N_1 = 1450$
----Si <u>tiramos de una vez 3000</u> monedas pueden salir entre 0 y 3000 caras (r). Es una B(3000 , 0,5) ; n=3000 ; N=1 Si obtenemos 1450 caras (c), habrá habido 1550 cruces (k). Como sólo se hace una vez, se suele asimilar al caso anterior y se dice que $N_0 = 1550$; $N_1 = 1450$, aunque realmente no es correcto. Mejor sería N_c y N_k
-----Si <u>tiramos tres monedas 1000 veces</u> y obtenemos 0 caras en 115 ocasiones, una cara en 380, dos caras en 370 y tres caras en 130: es una B(3 ; 0,5) , n=3 , N=1000 , N_0=115 , N_1=380 , N_2=370 y N_3=130

Problemas asociados a la DB

1) **calcular p(r)** : nos pueden pedir el cálculo de una r en concreto o de todas ellas.
Como ejemplo vemos la p de 2 caras lanzando 3 monedas. Es B(3 , 0,5)

1- aplicando la fórmula (de las dos que se han visto la más fácil es la segunda)

$$p(r=2)=\frac{3!}{2!*1!}0,5^2 0,5^1=0,3750$$

2- consultando la tabla (ver página 16) ya que en este caso se puede utilizar. Es una tabla de doble entrada con valores de n y r en la primera columna y ciertos valores de p en la primera fila.

En una B(3 , 0,5) p(r=2) = 0,3750

3- método intuitivo ("cuenta de la vieja"). Válido para una p elemental de 0,5.Veremos no sólo la p(r=3) sino todas las p(r). Hay que considerar todas las combinaciones posibles de cara (c) y cruz (k)

r	modalidades	$\binom{n}{r}$	p(r)
0	k k k	1	1/8
1	c k k k c k k k c	3	3/8
2	c c k c k c k c c	3	3/8
3	c c c	1	1/8
Σ		8	1

3/8 = 0,3750

2) **calcular N_r** : es decir, la frecuencia de cada modalidad al repetir el experimento binomial N veces $N_r = Np(r)$

Si el lanzamiento de las 3 monedas se repite 200 veces, teóricamente se obtendrán lo siguiente:

0 caras : N_0 = 200 * 1/8 = 25
1 cara : N_1 = 200 * 3/8 = 75
2 caras : N_2 = 200 * 3/8 = 75
3 caras : N_3 = 200 * 3/8 = 25

3)calcular la media, varianza, desviación estándar

$\bar{x} = np$; $s^2 = npq$; $s=\sqrt{npq}$

En el ejemplo de las monedas: $\bar{x}$ = 3 * 0,5 = 1,5
s^2 = 3 * 0,5 * 0,5 = 0,75
$s = \sqrt{3*0,5*0,5}$ = 0 ,866

4) **calcular los parámetros de una DB** , n y p, a partir de las frecuencias de las modalidades, es decir, a partir de Nr

n lo conocemos por los datos que nos dan.

p se calcula a partir de $\bar{x} = np$ y $\bar{x}=\frac{\sum(rN_r)}{N}$

Ejemplo:
Lanzadas 4 monedas 10000 veces se han obtenido los resultados que se muestran en la tabla: 0 caras en 4096 ocasiones, 1 cara en 4096, 2 caras en 1536, 3 caras en 256 y 4 caras en 16.

r	Nr	r*Nr
0	4096	0
1	4096	4096
2	1536	3072
3	256	768
4	16	64
∑	10000	8000

$$\overline{x} = \frac{8000}{10000} = 0'8 \qquad 0'8=4p \qquad p=0'2$$

por tanto es una B(4 , 0'2)

6) al crecer n la DB se llega a hacer inmanejable y la solución es **aproximarla** a otra Distribución fundamental transformando los parámetros originales en los propios de la distribución a la que se aproxima. Siempre que se cumplan ciertas condiciones.

- **a la DN ,** si p y q $\geq$ 0,1 (ó 10% si es %) y np y nq $\geq$ 5 (ó 10 y 500 si es un %) se verá al tratar la DN
- **a la DP ,** si p o q $\leq$ 0,1 (ó 10% si es %) y np o nq $\leq$ 5 (ó 10 y 500 si es %), aunque algunos admiten np o nq hasta 10 (ó 1000 si es %). Como veremos enseguida la DP es una variante de la DB y su parámetro λ es igual a n*p , por lo que la aproximación es muy fácil.

7) **comprobar el ajuste** de unos datos (una distribución real u observada) a una DB ideal
Para ello hay que calcular una distribución binomial teórica, que tenga los mismos parámetros que la real. Como partiremos de las frecuencias de cada modalidad, hay que utilizar el procedimiento visto en **5**). Luego se contrastan las frecuencias teóricas con las observadas por medio de una prueba de contraste de frecuencias, cuyo resultado se valora por χ^2. Si no se encuentran diferencias significativas, el ajuste es bueno, En caso contrario es malo.
Ejemplo: En un lote de 800 piezas cada una de las cuales tiene tres soldaduras se han observado las siguientes frecuencias de defectos de soldadura: 0 defectos en 97 ; 1 defecto en 305 ; 2 defectos en 297 y 3 defectos en 101. Comprobar el ajuste a una DB.

a) $\overline{x} = \dfrac{(0*97)+(1*305)+(2*297)+(3*101)}{800} = 1,5 \qquad p = \dfrac{1,5}{3} = 0,5$

b) cálculo de una B(3 ; 0,5) con N=800

r	p(r)	N_r
0	0,125	100
1	0,375	300
2	0,375	300
3	0,125	100
Σ		800

Las p (r) se pueden leer directamente en la tabla de la DB

recordar que $N_r = N*p(r)$

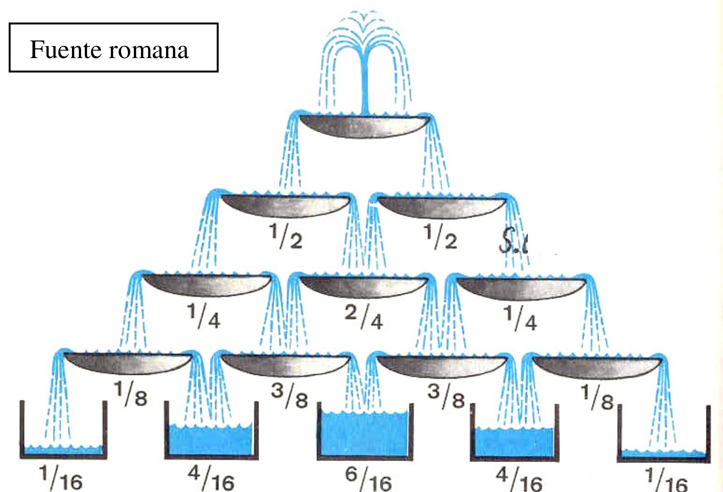
c) Ahora se contrastan las frecuencias observadas y las teórica:

f observadas	97	305	297	101
f teóricas	100	300	300	100

A simple vista se ve que el ajuste es muy bueno. Aplicando la prueba de contraste que veremos en el tema16 la z=0,213 que no es significativa y por tanto el ajuste es bueno.

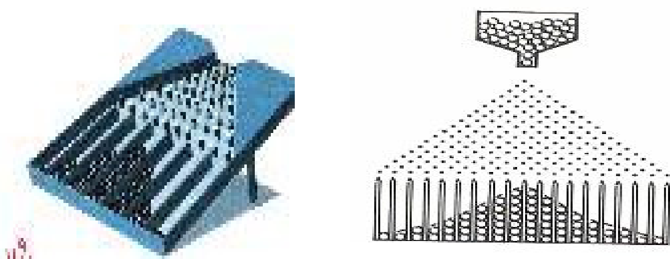
Modelos clásicos de la distribución binomial

Los más importantes son las fuentes romanas, el aparato de Galton y el triángulo de Pascal.



Fuente romana

La mitad del agua que sale por la fuente de arriba cae por cada lado. Y lo mismo ocurre con las demás fuentes. Al final unos recipientes recogen el agua. Siguiendo el camino del agua, se ve que el volumen recogido aumenta hacia en el centro. Una fuente perfecta sigue exactamente la DB. El primer recipiente corresponde a r=0, el 2º a r=1, el 3º a r=2 , etc El nº de recipientes por tanto es igual a n+1



El aparato de Galton sigue el mismo principio. Es una especie de embudo inclinado con filas de clavos, situados como las fuentes. Al final hay unos cajones receptores. Se lanza una bola que cada vez que choca con un clavo tiene la misma probabilidad de ir a la derecha que a la izquierda.

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1

El triángulo de Pascal empieza por el 1 de la primera fila. Los números de las otras filas se obtienen sumando los dos que están por encima de él a derecha e izquierda. Como en los lados siempre se suma el 1 con nada, todos son 1. Se pueden construir el nº de filas que uno quiera. En cada fila los números corresponden a los coeficientes $\binom{n}{r}$ para cada valor de r, de 0 a n. Por tanto n es igual al nº de coeficientes menos 1. La suma de los coeficientes de cada fila es igual a 2^n

DISTRIBUCIÓN DE POISSON

también llamada de los sucesos raros o de las probabilidades pequeñas.
Es una variante de la DB cuando p o q son muy pequeñas y n no es muy grande. En esta situación la DB se hace inexacta. La frontera se fija como se ha visto al tratar la aproximación de la DB a una DP en p ó q $\leq 0{,}1$ (ó el 10%, si se expresa en %; algunos admiten hasta 0,2 ó 20%) y np ó nq ≤ 5 (ó 500 si se expresa como %), aunque últimamente se acepta hasta 10 (ó 1000). Como en origen es una DB, es valido lo que hemos visto sobre n , r , N_r y N .
Aunque un suceso sea raro, ocurre de vez en cuando. Incluso con cierta frecuencia, si aumenta el nº de ocasiones para que ocurra. Ya vimos que la p de acertar 6 en la Primitiva es bajísima, pero como se hacen millones de apuestas, hay muchas semanas con uno o más acertantes. En un determinado cruce puede ser que la probabilidad de que un coche tenga un accidente sea muy baja, pero si el tráfico es muy intenso, puede haber accidentes incluso todos los días.
Al contrario, un hecho frecuente, como las llamadas que se reciben en la centralita telefónica de un hospital, se puede convertir en raro si consideramos las llamadas en una unidad de tiempo muy pequeña, p.e. segundos. En 24 horas quizá en la mayor parte de los segundos no haya ninguna llamada.
¡Fijarse también en q , no sólo en p! *. Una B(5 , 0'98) tiene la q=0,02 y debe ser aproximada a una P(4,9)*

Notación

P(λ) , siendo λ= np (λ es la letra griega lambda)

Cálculo de p(r)

$$p(r) = \frac{\lambda^r}{r!} e^{-\lambda}$$

el valor de $e^{-\lambda}$ (e es la base de los logaritmos neperianos) se puede hallar con una calculadora científica o leer en una tabla (página 15). La tabla tiene dos partes: una va de λ entre 0,00 y 0,99 . La otra parte da $e^{-\lambda}$ para valores enteros de λ entre 1 y 10. Para valores con decimales en este intervalo se descompone λ en dos partes: una entera y la otra decimal . Por ejemplo: λ = 3,48 se descompone en 3 y 0,48. Los valores de $e^{-\lambda}$ se pueden leer en la tabla y hay que multiplicarlos, ya que este procedimiento se basa en que el producto de dos potencias de la misma base es otra potencia con la mima base y cuyo exponente es la suma de los exponentes.
Ejemplos: Calcular p(r=3) para una P(0,25) y para una P(3,48)

1) $p(r=3) = \frac{0,25^3}{3!} e^{-0,25} = 0,0020$

2) $p(r=3) = \frac{3,48^3}{3!} e^{-3,48} = 7,024 * (0,04979 * 0,6188) = 0,2164$

Media, varianza y desviación estándar

$\bar{X} = \lambda = np$ $\bar{X} = \frac{\sum (rN_r)}{N}$ $s^2 = \lambda$ $s = \sqrt{\lambda}$

Gráfico : es también el diagrama de barras

Problemas asociados a la DP

son similares a los vistos en la DB, ya que es una variante de la misma.
1) **calcular p(r)** : utilizando la fórmula

2) **calcular N_r** : es decir, la frecuencia de cada modalidad al repetir el experimento N veces

$$N_r = N * p(r)$$

3) **calcular el parámetro λ** a partir de las frecuencias de las modalidades, es decir, a partir de Nr , **utilizando las fórmulas ya conocidas de la DB :** $\overline{X} = np$, $\overline{X} = \frac{\sum(rN_r)}{N}$ y **λ = np**

4) **calcular la media, varianza, desviación estándar :** $\overline{X} = \lambda = np$; $s = \sqrt{\lambda}$; $s^2 = \lambda$

5) **comprobar el ajuste** de unos datos a una DP

Veremos un ejemplo para comprobar el ajuste de una distribución real a una DP teórica. Sabemos que a partir de los datos que nos den hay que calcular el parámetro λ. Luego se calculan las p teóricas asociadas a cada una de las modalidades deseadas y se multiplican por N, obteniendo de esta forma las N_r teóricas, que hay que contrastar con las observadas mediante la prueba estadística correspondiente.
--El veterinario militar alemán Borotkiewitz estudió las defunciones por coces de caballo en 20 regimientos prusianos durante 10 años("Ley de los pequeños números", 1898). Encontró que seguían la distribución de los sucesos raros de Poisson y que por tanto eran fruto del azar y no eran imputables en principio a fallos de organización.
De los 200 regimientos-año (20*10) hubo 109 que no registraron muertes, 65 con un fallecimiento, 22 con dos, 3 con tres y 1 con cuatro.

Como λ es igual a la media, se utiliza la fórmula ya conocida $\overline{X} = \frac{\sum(rN_r)}{N}$

r	N_r
0	109
1	65
2	22
3	3
4	1
Σ	200

$$\overline{x} = \frac{(0*109)+(1*65)+(2*22)+(3*3)+(4*1)}{200} = 0{,}61$$

Hay que desarrollar una P(0,61) con N=200

r	p(r)	N_r
0	0,543	109
1	0,331	66
2	0,101	20
3	0,021	4
4	0,003	1
Σ		200

Los valores de N_r se presentan redondeados para que se vea mejor a simple vista la comparación con los observados. Para el contraste con las frecuencias observadas habría que dejar dos o tres decimales (esto es válido para cualquier ajuste).
La prueba da z=0,465 que no es significativa.
Por tanto el ajuste de esos datos a una DP es bueno

DISTRIBUCION NORMAL

Es la distribución típica de variables aleatorias cuantitativas continuas cuando el tamaño es grande (por consenso, cuando N≥30) . Sus parámetros básicos son la media y la desviación estándar. Su desarrollo se debe fundamentalmente a Laplace y Gauss. Quetelet le dió el nombre de normal o natural porque observó que la gran mayoría de variables fisiológicas seguían este modelo. Es un nombre consagrado por el uso y no quiere decir que las otras distribuciones sean "anormales". Los norteamericanos usan y han exportado la denominación de "distribución gaussiana".
Siguen la DN todo tipo de variables biológicas (como frecuencia cardíaca, tensión arterial, componentes químicos de la sangre y orina, medidas corporales...), duración o vida de objetos y seres vivos, etc

Notación : $N(\bar{x}, s)$

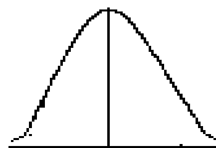
Fórmula

La fórmula para calcular las p asociadas a intervalos de valores (no se pueden calcular p de valores puntuales, ya que en el contexto de la DN son infinitésimos) es muy compleja y necesita integración. Pero afortunadamente no hay que utilizarla, pues se dispone de una tabla de fácil manejo, que nos da el cálculo ya hecho. A título informativo la fórmula es:

$$p(a \le x \le b) = \int_a^b f(x)\, d(x) \quad , \text{siendo } dx = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{s}\right)^2}$$

Representación gráfica

es la curva o campana de Gauss, en “chapeau de gendarme” (gorro de gendarme) de los tiempos napoleónicos. Es el límite de un histograma cuando la amplitud de las clase se hace infinitesimal y el nº de datos tiende a infinito.



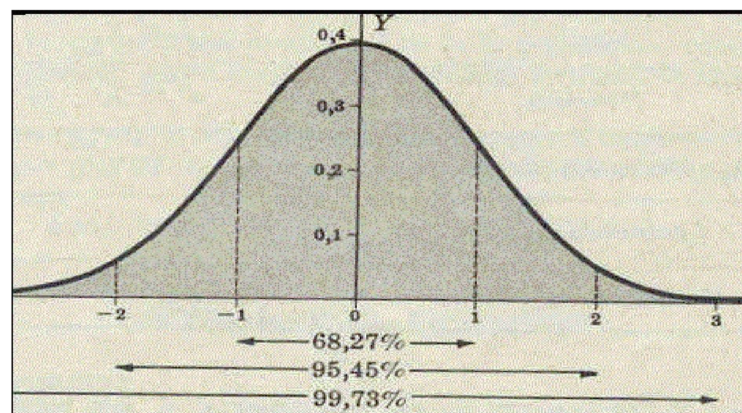
Es **simétrica** alrededor de un eje vertical que pasa por $\bar{x}$ y **asintótica** al eje de abscisas (lo corta en el infinito por ambos lados, aunque a partir de $\bar{x} \pm 3s$ ya casi lo toca). La campana engloba todos los valores y por tanto la p de que un valor cualquiera esté en ella es 1 ó 100%. La superficie de campana delimitada por dos valores del eje de abscisas equivale a la probabilidad de que un valor cualquiera se encuentre en ese área. Cada distribución tiene su propia campana, hay infinitas curvas de DN. En estas condiciones su manejo sería muy difícil y complicado, ya que habría que aplicar cada vez la fórmula. Afortunadamente se ha encontrado un modelo único de distribución y por tanto de campana al que pueden ser adaptadas todas las DN. Es la llamada DN tipificada.

Tipificación

Consiste en transformar cualquier $N(\bar{x}, s)$ en otra $N(0, 1)$, es decir, en una DN de media 0 y desviación estándar 1. Para ello hay que transformar los valores originales x en puntuaciones estándar o valores tipificados, que aquí llamaremos c. (Otros nombres: z o SDS).

$$c = \frac{x - \bar{x}}{s}$$

Entre dos valores de c quedan delimitadas áreas (=probabilidad) que se pueden obtener a partir de la **tabla de la DN** tipificada. Ya se ha dicho al principio que no se pueden calcular p de valores aislados, sólo de intervalos más o menos grandes.



En esta campana están representadas las áreas o probabilidades entre valores de c +1 y –1 , +2 y –2 , +3 y –3 . Pero es preferible expresar la p con números más "redondos" :
---Al intervalo entre c = -1,96 y c = 1,96 corresponde un 95% de la superficie de la campana.

$p(-1{,}96 \leq c \leq 1{,}96) = 0{,}95$ ó 95%

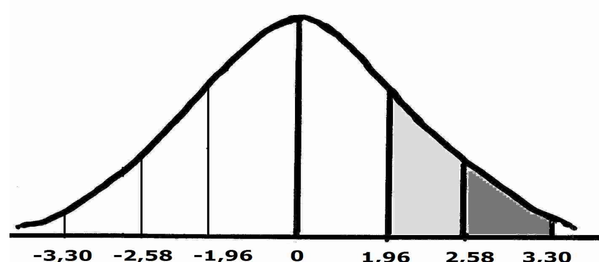
---Al intervalo entre c = -2,58 y c = 2,58 corresponde un 99% de la superficie de la campana.

$p(-2{,}58 \leq c \leq 2{,}58) = 0{,}99$ ó 99%

---Al intervalo entre c = -3,30 y c = 3,30 corresponde un 99,9% de la superficie de la campana.

$p(-3{,}30 \leq c \leq 3{,}30) = 0{,}999$ ó 99,9%

que son los que utilizaremos aquí.



Es imprescindible dibujar una campana y marcar en ella la media y el valor o valores de x.
Una vez tipificada se anotan el los valores de c. A la media le corresponde siempre por definición el valor de 0.

Tabla de la DN tipificada

El modelo que utilizamos es de media campana, va de 0 a + ∞. (Página 16). Hay otro con la campana entera, que abarca de - ∞ a + ∞. Nos da la p de que un valor cualquiera esté entre c = 0 y otro valor de c. Al ser la campana simétrica sirve por igual para valores de c positivos o negativos, siempre con dos decimales. Es una tabla de doble entrada. En la primera columna están valores de c con un decimal y en la primera fila está el segundo decimal. Donde confluyen ambos está la probabilidad buscada.

Problemas asociados a la DN

1---**tipificar**
p.e. x=5 y x=3 de una B(4 , 2)

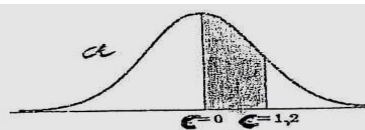
→ c = (5-4)/2 = 0,5 → c = (3-4)/2 = -0,5

2---**calcular la probabilidad de un intervalo**,
p.e. entre c = 0 y c = 0,46

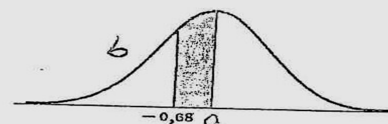
→ $p(0 \leq c \leq 0{,}46) = 0{,}1772$

CASOS POSIBLES

a) $p(0 \leq c \leq 1{,}2) =$ 0,3849 ó 38,5%



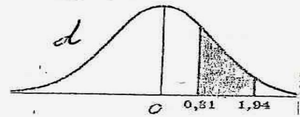
b) $p(-0{,}68 \leq c \leq 0) =$ 0,2518 ó 25,2%



c) $p(-0{,}46 \leq c \leq 2{,}21) =$ 0,6636 ó 66,4%
=área entre –0,46 y 0 más área entre 0 y 2,21

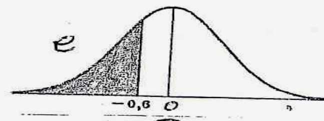


d) $p(0{,}81 \leq c \leq 1{,}94) =$
$0{,}1828$ ó $18{,}3\%$



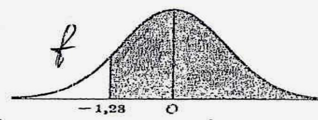
=área para c=1,94 menos
área para c=0,81

e) $p(c \leq -0{,}6) = 0{,}2742$ ó $27{,}4\%$



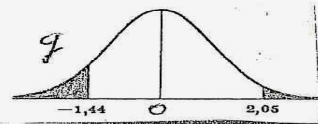
= 0,5 menos área para c=-0,6

f) $p(c \geq -1{,}28) = 0{,}8997$ ó 90%



= 0,5 más área para c=-1,28

g) $p(c \leq 1{,}44$ y $c \geq 2{,}05) =$
$0{,}0951$ ó $9{,}5\%$

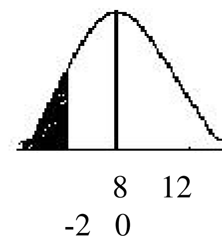


= 1 – área para c=-1,44 y c=2,05

Ejemplo:
La duración media de una bombilla es de 12 meses, con una varianza de 4. El fabricante garantiza que dura más de 8 meses. Calcular
1) la probabilidad de que se funda en el periodo de garantía
2) la probabilidad de que dure al menos 16 meses
3) la probabilidad de que dure entre 15 y 18 meses

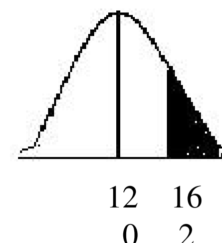
La variable “Vida de la bombilla” es una N(12, 2)

1) $p(x \leq 8)$?



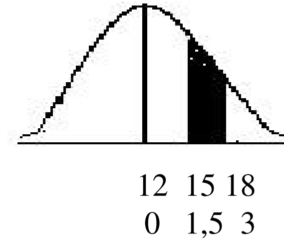
se dibuja la campana
se tipifica: $c = (8-12)/2 = -2$
$p(c \leq -2) = 0{,}5 - p(-2 \leq c \leq 0) =$
$0{,}5 - 0{,}4772 = 0{,}0228$ ó $2{,}28\%$

2) $p(x \geq 16)$?



$c = (16-12)/2 = 2$
$p(c \geq 2) = 0{,}5 - p(0 \leq c \leq 2) =$
$0{,}5 - 0{,}4772 = 0{,}0228$ ó $2{,}28\%$

3) $p(15 \le x \le 18)$?



$c_1 = (15-12)/2 = 1,5$
$c_2 = (18-12)/2 = 3$

$p(1,5 \le c \le 3) = p(0 \le c \le 3) - p(0 \le c \le 1,5)$
$= 0,4987 - 0,4332 = 0,0655$ ó 6,55%

3)---**calcular la frecuencia de un intervalo**, conocidos N y la p del intervalo.
Es similar a lo visto en la DB: $N_r = N * p$. Aquí para simplificar llamaremos al intervalo i (en vez de $a \le x \le b$ ó $\in (a \div b)$) y a su frecuencia Ni.
Supongamos que en una muestra de 6500 individuos en los que se hecho el análisis A hemos calculado una p de 0,2426 para el intervalo entre 7 y 10 mg/dl. ¿Cuantos individuos tendrán ese análisis entre 7 y 10 mg/dl?
Solución: $Ni = 6500 * 0,2426 = 1576,9 \approx 1577$

4)---**Calcular un valor de c a partir de una p y de un punto de referencia en la campana** (es decir, de otro valor de c)
Como en todos los problemas de campana es imprescindible dibujarla y situar en ella el punto c de referencia.
No olvidar que los de signo positivo se ponen a la derecha de la media (según vemos la campana) y los negativos a la izquierda.
Luego se busca en la tabla la p que nos dan y se ve a que valor de c corresponde. No olvidar el signo menos si le corresponde estar a la izquierda. Si el valor de p no está exactamente se toma el más próximo, siguiendo el mismo procedimiento que en el redondeo.

CALCULO DE UN VALOR c A PARTIR DE UNA PROBABILIDAD Y UN PUNTO DE REFERENCIA.

1) el área entre 0 y c es de 0.3770 ; $p \in (0 \div c) = 0.3770$

- dibujar campana

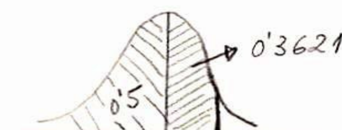


- buscar en la tabla. Vemos que le corresponde un c de 1.16

respuestas: hay dos c= 1.16 y c= -1.16

2) el área a la izquierda de c es 0.8621 ; $0.8621 = p \in (-\infty \div c)$

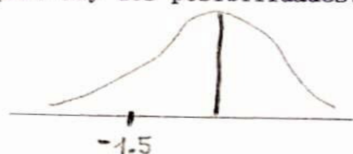
- dibujar la campana; al se p>0.5 c tiene que estar en el lado derecho



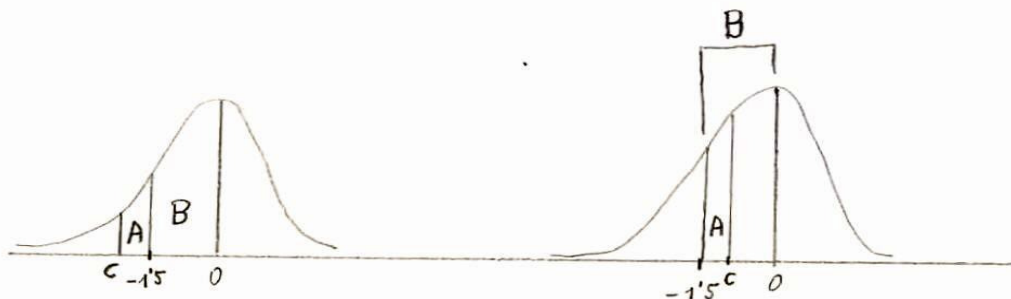
- como nuestra tabla es de sólo media campana restamos 0.5 p=0.3621
- buscamos en la tabla y encontramos una c de 1.09
respuesta: 1.09

3) el área entre −1.5 y c es de 0.0217 ; 0.0217 = p ∈(−1.5 ÷ c)

- dibujar campana; c tiene que estar por fuerza a la izquierda ya que si fuera + el área valdría más de 0.4332 que es la p que corresponde a c=−1.5 pero hay dos posibilidades: a la derecha y a la izquierda de −1.5



-- p = A+B = 0.0217+0.4332 = 0.4549 ; le corresponde c ≈ −1.69
-- p = B−A = 0.4332−0.0217 = 0.4115 ; le corresponde c ≈ −1.35



5)---**Calcular una puntuación original, x , a partir de puntuaciones estándar c**

Se utiliza la fórmula $\mathbf{c = \frac{x - \bar{x}}{s}}$; puede ser necesario dibujar la campana si hay alguna duda.

Ejemplos:

a)—Calcular la puntuación original que corresponde a una c = 1,6 en una N(6 , 2)

→ 1,6 = (x-6)/2 ; x = 9,2

b)—En esa misma distribución calcular la puntuación original que deja por debajo de ella el 86,21% de los valores.

→ 86,21% equivale a una p de 0,8621 , por lo que x tiene que estar situado en el lado derecho de la campana. Para poder utilizar la tabla le restamos 0,5 a 0,8621 y queda 0,3621 . Le corresponde una c = 1,09 . Entonces 1,09 = (x-6)/2 ; x = 8,18

6)—**Calcular $\bar{x}$ y s** a partir de otros parámetros.

Se utiliza la misma fórmula: $\mathbf{c = \frac{x - \bar{x}}{s}}$.

De sus 4 elementos hay que conocer 3. Puede ser conveniente dibujar la campana.

Ejemplo: Calcular la s de un DN cuya media es 5 y en la que p(x≤6) = 0,6064

→ x tiene que esta en el lado derecho de la campana al ser la p > 0,5

0,6064 – 0,5 = 0,1064 a quien corresponde una c de 0,27 .

0,27 = (6-5)/s   y   s = 3,70

7)---**aproximar una DB o una DP a una DN**

Ambas se aproximan de forma perfecta a la DN cuando np ó λ → ∞ .

Las condiciones para la aproximación de la DN de una DB, recordemos, son p y q ≥ 0,1 (ó 10%) y np y nq ≥ 5 (ó 500, si p se expresa como %).

La DB se transforma en una DN, que tenga la misma media y desviación estándar que la DB

La DP se aproxima de forma similar.

Hay que hacer una pequeña corrección, la llamada **corrección de continuidad**. La DB es discreta y por tanto discontinua y la DN es continua. No se toman los límites tabulados del intervalo sino el límite real que corresponda. Los límites tabulados deben quedar incluidos, por lo que en unos casos se tomará el límite real inferior y en otros el superior.
Así, si tiramos 300 monedas y queremos saber la p de obtener entre 90 y 120 caras, no calcularemos $p(90 \leq x \leq 120)$ sino $p(89{,}5 \leq x \leq 120{,}5)$.

Ejemplo: Esta misma tirada de las 300 monedas. Es una B(300 , 0,5). $\bar{x} = 300 * 0{,}5 = 150$
$s = \sqrt{npq} = 8{,}66$. Por tanto la transformamos en un N(150 , 8,66) , en la que hay que calcular $p(89{,}5 \leq x \leq 120{,}5)$ por el procedimiento ya visto.(Es como el caso 2d, pero en el lado izquierdo de la campana. El resultado es 0,0003)

8)---**Comprobar el ajuste de una distribución real (observada) a una DN**.
Lo veremos con la distribución de la talla de sus compañeros del curso 1978/79.
N = 47 $\bar{x} = 167{,}9$ cm s = 7,8 cm

Talla de los alumnos de Bioestadística Curso 1978/79		
clases	p.m.	nº
152-161 cm	156,5	10
162-171 cm	166,5	23
172-181 cm	176,5	12
182-191 cm	186,5	2

Hay que construir una DN teórica que tenga los mismos parámetros que la real. Una vez conocidas las frecuencias teóricas de cada clase se contrastan con las reales, mediante la prueba correspondiente. Si no hay diferencias significativas, el ajuste es bueno.

El procedimiento es un tanto engorroso y conviene seguir una metódica clara para no equivocarse. Como la que se usa aquí.

Pasos:
1) construirse una tabla auxiliar

2) comenzar a rellenarla por los Límites Reales

clases	L. reales	c	área entre c y 0	p de la clase	Ni teórico ≈		Ni real
	- ∞	- ∞					
	151,5						
	161,5						
	171,5						
	181,5						
	191,5						
	+ ∞	+ ∞					

3) situar las clases

clases	L. reales	c	área Entre c y 0	p de la clase	Ni teórico ≈		Ni real
	- ∞						

	151,5						
152-161							
	161,5						
162-171							
	171,5						
172-181							
	181,5						
182-191							
	191,5						

	+ ∞						

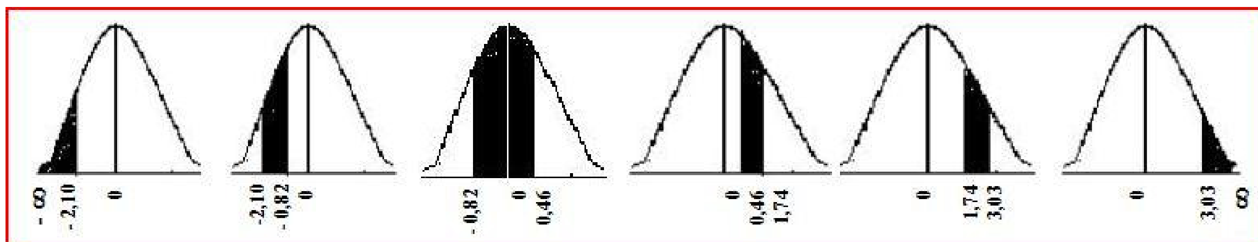
4) Calcular valores de c para cada L. real y el área entre c y 0

clases	L. reales	c	área entre c y 0	p de la clase	Ni teórico		Ni real
	- ∞	- ∞	0,5				

	151,5	-2,10	0,4821				
152-161							
	161,5	-0,82	0,2939				
162-171							
	171,5	0,46	0,1772				
172-181							
	181,5	1,74	0,4591				
182-191							
	191,5	3,03	0,4988				

	+ ∞	+ ∞	0,5				

5) calcular la p de cada clase (dibujar campana), pasarla a la tabla auxiliar y calcular Nr teórico



clases	L. reales	c	Área (p) entre c y 0	p de la clase	Ni teórico ≈		Ni real
	- ∞	- ∞	0,5				
--------------				0,0179	0,9		--
	151,5	-2,10	0,4821				
152-161				0,1882	9		10
	161,5	-0,82	0,2939				
162-171				0,4711	22		23
	171,5	0,46	0,1772				
172-181				0,2819	13		12
	181,5	1,74	0,4591				
182-191				0,0397	2		2
	191,5	3,03	0,4988				
--------------				0,0012	0,1		--
	+ ∞	+ ∞	0,5				

6) aplicar prueba de contraste de frecuencias (fórmula nº 3; tema 16). Se obtiene Z=1,233, que es < $\chi 2$ (5 , 0'05)=11,07, n.s. Se concluye que el ajuste es bueno, como parece ya a simple vista.

Distribución de la t de Student

es la distribución teórica de las muestras pequeñas de una población que sigue la ley normal con datos cuantitativos continuos.
Gosset (que utilizaba el seudónimo de Student) comprobó que cuando disminuía el tamaño de las muestras, no valían del todo los normas de la DN, tanto más cuanto más pequeña sea la muestra. Hasta N=30 las diferencias son bastante acusadas. Por eso la mayoría de autores ponen a ese nivel la frontera de uso práctico entre DN y t de Student.. Otros lo ponen en 60 y algunos hasta en 120. Los programas estadísticos utilizan casi exclusivamente la t de Student para todas las variables continuas, ya que hasta el infinito no se produce una identidad plena entre ambas distribuciones. La DN está en vías de extinción, al menos en la práctica. Nosotros seguiremos el criterio de utilizar la t de Student para muestras pequeñas (N<30) y la DN para las grandes.

La **notación** es t(gl , α). α es el nivel de significación elegido y gl es el grado de libertad. Con este nombre se designa al número de observaciones independientes, que en general son N-1. Un ejemplo ayudará a entender este concepto. Si nos piden 5 valores que sumen 35, sólo podremos elegir libremente 4, pues el 5º es obligado: supongamos que elegimos 8 , 10 , 23 , -15 . El 5º número tiene que ser por fuerza 9 ; hay 4 grados de libertad.

Aquí no hay modelo tipificado y para cada valor de N hay una campana distinta (que no es preciso dibujar..).

La **tabla** sigue el modelo de las tablas de doble entrada. En la primera columna está el grado de libertad y en la primera fila hay tres niveles de significación.
t(5 , 0,05) = 2,571 ; t(26, 0,001) = 3,707 ; t(15, 0,01) = 2,947

El término t se usa para designar varias cosas, lo que puede generar cierta confusión:
1—la distribución de la t de Student
2---los valores de la abscisa de la campana correspondiente, donde están los valores de referencia para valorar el resultado de las pruebas. Es el equivalente a la c de la DN
3---el resultado de las pruebas estadísticas que son valoradas por la t de Student. Esto lo obviamos llamando de una forma genérica **Z** a todos los resultados de las pruebas estadísticas, nombre arbitrario que puede ser sustituido por cualquier otro.

Distribución χ^2 (chi o ji cuadrado)

es la distribución que siguen las frecuencias de muestras obtenidas de una población.
También aquí hay grados de libertad y para cada grado de libertad hay un gráfico distinto.
Notación: χ^2 (gl, α)
La **tabla** es también de doble entrada, con una disposición similar, aunque nos ofrece un nivel de significación más, el de 0,02.
χ^2(1 , 0,05) = 3,84 ; χ^2 (2 , 0,01) = 9,21 ; χ^2 (5 , 0,001) = 20,52
Su uso es típico de las tablas de 2 por 2 (2x2) ó f por k (fxk),siendo f el nº de filas y k el de columnas.
Con el nombre de χ^2 se pueden designar también dos cosas:
1---la distribución χ^2
2---los resultados de las pruebas que son valoradas por la χ^2 (lo que no seguimos aquí, pues a todos los resultados los llamamos Z, con independencia de cómo sean valorados).

Distribución de la F de Snedecor-Fisher

es la distribución de los posibles cocientes de dos varianzas, poniendo siempre la mayor de ellas en el numerador. Así F será siempre ≥ 1, lo que supone un ahorro de espacio al confeccionar la tabla. Aquí también hay grados de libertad y gráficos distintos para cada grado de libertad (que no tenemos que dibujar).
Notación : F(gl1 , gl2 , α) . Siendo gl1 = k-1 (k es el nº de muestras o grupos) y gl2 = (N-1)(k-1).
N es la frecuencia total, el tamaño total de todas las muestras o grupos .
Tablas: para cada nivel de significación hay una tabla distinta, que también es de doble entrada.
Se busca gl1 en la primera fila y gl2 en la primera columna.
F(5 , 9 , 0,05) = 3,48 ; F(12 , 10 , 0,01) = 4,71
Cuando la tabla no nos ofrece el valor exacto del gl, se aproxima al más cercano o si se es muy riguroso, siempre al inferior. Para F(90 , 30 , 0,001) lo habitual es elegir 2,76, pero en función del rigor de la investigación se puede elegir también.2,92
Se usa para valorar la llamada “igualdad de varianzas” y los resultados de las pruebas de ANOVA.
Con F se pueden designar también dos cosas:
1---la distribución F
2---los resultados de las pruebas que son valoradas por la F (lo que no seguimos aquí, pues a todos los resultados los llamamos Z, con independencia de cómo sean valorados).

DISTRIBUCIÓN HIPERGEOMETRICA

Variante de Binomial cuando no hay reposición de efectivos y N es finita. Si N es muy grande, vale la Binomial.(La aproximación es ya buena, si $N_1/N \leq 0{,}1$ ó mejor si $\leq 0{,}05$). O sea, siempre que el tamaño de la muestra sea el 10% o menos del tamaño de la población, se puede usar –y de hecho se usa- la DB.

Notación: $H(n , N , N_1)$, siendo n como en la DB, N el nº total de individuos y N_1 los que presentan la característica. Se busca la p de r (que va de 0 a n, como en la DB).

Fórmula: $$\mathbf{p(r)} = \frac{\binom{N1}{r}\binom{N-N1}{n-r}}{\binom{N}{n}} = \frac{\dfrac{N1!}{r!(N1-r)!} * \dfrac{(N-N1)!}{(n-r)!(N-N1-n+r)!}}{\dfrac{N!}{n!(N-n)!}}$$

ó simplificando : $$\mathbf{p(r)} = \frac{N1!(N-N1)!n!((N-n)!}{r!(N1-r)!(n-r)!(N-N1-n+r)!N!}$$

La varianza es menor que en la DB: $$s^2 = npq\frac{N-n}{N-1}$$

Al intervenir tantas factoriales en la fórmula, las calculadoras e incluso muchos programas estadísticos de ordenador se ven sobrepasados fácilmente en su capacidad de cálculo. La hoja de cálculo Excel admite hasta N = 170, mientras otros programas más antiguos, basados en MS-Dos, no pasan de 33. Lo vemos aquí para completar el tema, ya que por este motivo no puede ser objeto de examen. En la práctica es habitual hacer los cálculos como si fuera una DB, ya que el error es en general muy pequeño.

Ejemplo 1: De 100 enfermos, 20 presentan una infección. Se toman 5 al azar y se pide la probabilidad de que sólo 1 presente la infección.
$N=100$; $N_1=20$; $n=5$; $r=1$ Es H(5 , 100 , 20)
Haciendo las operaciones sale p(r=1) = **0,420144**...
Como binomial sería B(5 , 0.2) y p(r=1)=0,4096

Ejemplo 2:
p de que sacando 4 cartas de una baraja española de 40 cartas, las 4 sean ases.
Es H(4 , 40 , 4) y p(r=4) = $\mathbf{1{,}0942*10^{-5}}$

Como B(4, 4/40) = B(4 , 0.1) , p(r=4) = 0,0001

Por cálculo elemental (que es exacto) : 4/40 * 3/39 * 2/38 * 1/37 = 24/2193369 = $\mathbf{1{,}0942*10^{-5}}$

Por Poisson , P(0,4) : p(r=4) = 0,0007

Distribución Binomial B(n , p)

$$\bar{X} = np \qquad s=\sqrt{npq} \qquad N=\sum N_r \qquad N_r = Np(r)$$

$$\bar{X} = \frac{\sum (rN_r)}{N} \qquad p(r)=\frac{n!}{r!*(n-r)!}p^r q^{(n-r)}$$

n	r	p 0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039

Distribución de Poisson $\quad P(\lambda)$

$$p(r) = \frac{\lambda^r}{r!} e^{-\lambda}$$

$\bar{X} = \lambda = np \qquad \bar{X} = \frac{\sum(rN_r)}{N} \qquad s^2 = \lambda \qquad s = \sqrt{\lambda}$

Valores de $e^{-\lambda}$

(De *Statistics*, por M. R. Spiegel. Schaum Publishing Company. Nueva York, 1961.)

λ	0	1	2	3	4	5	6	7	8	9
0,0	1,0000	0,9900	0,9802	0,9704	0,9608	0,9512	0,9418	0,9324	0,9231	0,9139
0,1	0,9048	0,8958	0,8869	0,8781	0,8694	0,8607	0,8521	0,8437	0,8353	0,8270
0,2	0,8187	0,8106	0,8025	0,7945	0,7866	0,7788	0,7711	0,7634	0,7558	0,7483
0,3	0,7408	0,7334	0,7261	0,7189	0,7118	0,7047	0,6977	0,6907	0,6839	0,6771
0,4	0,6703	0,6636	0,6570	0,6505	0,6440	0,6376	0,6313	0,6250	0,6188	0,6126
0,5	0,6065	0,6005	0,5945	0,5886	0,5827	0,5770	0,5712	0,5655	0,5599	0,5543
0,6	0,5488	0,5434	0,5379	0,5326	0,5273	0,5220	0,5169	0,5117	0,5066	0,5016
0,7	0,4966	0,4916	0,4868	0,4819	0,4771	0,4724	0,4677	0,4630	0,4584	0,4538
0,8	0,4493	0,4449	0,4404	0,4360	0,4317	0,4274	0,4232	0,4190	0,4148	0,4107
0,9	0,4066	0,4025	0,3985	0,3946	0,3906	0,3867	0,3829	0,3791	0,3753	0,3716

(λ = 1, 2, 3, ..., 10)

λ	1	2	3	4	5	6	7	8	9	10
$e^{-\lambda}$	0,36788	0,13534	0,04979	0,01832	0,006738	0,002479	0,000912	0,000335	0,000123	0,000045

NOTA. Para obtener valores de $e^{-\lambda}$ para otros valores de λ basta tener en cuenta las reglas del producto de potencias, por ejemplo:

$$e^{-3,48} = e^{-3,00} \cdot e^{-0,48} = 0,04979 \cdot 0,6188 = 0,03081.$$

ejemplos:

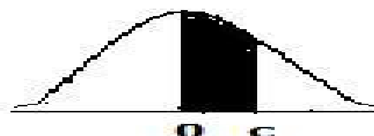
$e^{-0,28} = 0,7558 \qquad e^{-0,95} = 0,3867$

$e^{-0,50} = 0,6065 \qquad e^{-5} = 0,001832$

$e^{-3,48} = e^{-3} * e^{-0,48} = 0,04979 * 0,6188 = 0,03081$

Distribución normal N (0 , 1)

$$c = \frac{x - \bar{x}}{s}$$



la tabla da la probabilidad de que un valor cualquiera esté entre c = 0 y otro valor de c

a esta c se la llama hoy día mayoritariamente **z**

c	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	**0,1915**	**0,1950**	**0,1985**	**0,2019**	**0,2054**	**0,2088**	**0,2123**	**0,2157**	**0,2190**	**0,2224**
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	**0,3413**	**0,3438**	**0,3461**	**0,3485**	**0,3508**	**0,3531**	**0,3554**	**0,3577**	**0,3599**	**0,3621**
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	**0,4332**	**0,4345**	**0,4357**	**0,4370**	**0,4382**	**0,4394**	**0,4406**	**0,4418**	**0,4429**	**0,4441**
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	**0,4772**	**0,4778**	**0,4783**	**0,4788**	**0,4793**	**0,4798**	**0,4803**	**0,4808**	**0,4812**	**0,4817**
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	**0,4938**	**0,4940**	**0,4941**	**0,4943**	**0,4945**	**0,4946**	**0,4948**	**0,4949**	**0,4951**	**0,4952**
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	**0,4987**	**0,4987**	**0,4987**	**0,4988**	**0,4988**	**0,4989**	**0,4989**	**0,4989**	**0,4990**	**0,4990**
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Tabla de $\chi 2$

p

g. l.	**0,05**	**0,02**	**0,01**	**0,001**
1	3,84	5,41	6,64	10,83
2	5,99	7,82	9,21	13,82
3	7,81	9,84	11,34	16,27
4	9,49	11,77	13,28	18,47
5	11,07	13,39	15,09	20,52
6	12,59	15,03	16,81	22,46
7	14,07	16,62	18,48	24,32
8	15,51	18,17	20,09	26,13
9	16,92	19,68	21,67	27,88
10	18,31	21,16	23,21	29,59

Tabla de la t de Student

p *p*

g. l.	**0,05**	**0,01**	**0,001**		**g. l.**	**0,05**	**0,01**	**0,001**
1	12,71	63,66	636,6		26	2,056	2,779	3,707
2	4,303	9,925	31,60		27	2,052	2,771	3,690
3	3,182	5,841	12,94		28	2,048	2,763	3,674
4	2,776	4,604	8,610		29	2,045	2,756	3,659
5	2,571	4,032	6,859		30	2,042	2,750	3,646
6	2,447	3,707	5,959		35	2,030	2,724	3,592
7	2,365	3,499	5,405		40	2,021	2,704	3,551
8	2,306	3,355	5,041		45	2,014	2,689	3,521
9	2,262	3,250	4,781		50	2,008	2,678	3,496
10	2,228	3,169	4,587		55	2,004	2,669	3,476
11	2,201	3,106	4,437		60	2,000	2,660	3,460
12	2,179	3,055	4,318		70	1,994	2,648	3,435
13	2,160	3,012	4,221		80	1,989	2,638	3,416
14	2,145	2,977	4,140		90	1,986	2,631	3,402
15	2,131	2,947	4,073		100	1,982	2,626	3,390
16	2,120	2,921	4,015		120	1,980	2,617	3,373
17	2,110	2,898	3,965		130	1,977	2,612	3,361
18	2,101	2,878	3,922		140	1,975	2,607	3,352
19	2,093	2,861	3,883		150	1,974	2,605	3,349
20	2,086	2,845	3,850		160	1,973	2,603	3,346
21	2,080	2,831	3,819		200	1,972	2,601	3,340
22	2,074	2,819	3,792		300	1,968	2,592	3,340
23	2,069	2,807	3,767		400	1,966	2,588	3,315
24	2,064	2,797	3,745		500	1,965	2,586	3,310
25	2,060	2,787	3,725		∞	1,960	2,576	3,291

F de Snedecor-Fisher $\alpha = 0,05$

g.l. 1 →

g.l. 2 ↓	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500	∞
5	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,56	4,52	4,50	4,48	4,46	4,44	4,41	4,37	4,36
6	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,87	3,83	3,81	3,79	3,77	3,75	3,71	3,68	3,67
7	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51	3,44	3,40	3,38	3,36	3,34	3,32	3,27	3,24	3,23
8	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22	3,15	3,11	3,08	3,06	3,04	3,02	2,97	2,94	2,93
9	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01	2,94	2,89	2,86	2,84	2,83	2,80	2,76	2,72	2,71
10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85	2,77	2,73	2,70	2,68	2,66	2,64	2,59	2,55	2,54
11	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72	2,65	2,60	2,57	2,55	2,53	2,51	2,46	2,42	2,40
12	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62	2,54	2,50	2,47	2,44	2,43	2,40	2,35	2,31	2,30
13	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53	2,46	2,41	2,38	2,36	2,34	2,31	2,26	2,22	2,21
14	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46	2,39	2,34	2,31	2,28	2,27	2,24	2,19	2,14	2,13
15	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40	2,33	2,28	2,25	2,22	2,20	2,18	2,12	2,08	2,07
16	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35	2,28	2,23	2,19	2,17	2,15	2,12	2,07	2,02	2,01
17	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31	2,23	2,18	2,15	2,12	2,10	2,08	2,02	1,97	1,96
18	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27	2,19	2,14	2,11	2,08	2,06	2,04	1,98	1,93	1,92
19	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,16	2,11	2,07	2,05	2,03	2,00	1,94	1,89	1,88
20	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20	2,12	2,07	2,04	2,01	1,99	1,97	1,91	1,86	1,84
21	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,22	2,20	2,18	2,10	2,05	2,01	1,98	1,96	1,94	1,88	1,83	1,81
22	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,20	2,17	2,15	2,07	2,02	1,98	1,96	1,94	1,91	1,85	1,80	1,78
23	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,18	2,15	2,13	2,05	2,00	1,96	1,93	1,91	1,88	1,82	1,77	1,76
24	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,15	2,13	2,11	2,03	1,97	1,94	1,91	1,89	1,86	1,80	1,75	1,73
25	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,11	2,09	2,01	1,96	1,92	1,89	1,87	1,84	1,78	1,73	1,71
26	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,12	2,09	2,07	1,99	1,94	1,90	1,87	1,85	1,82	1,76	1,71	1,69
27	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13	2,10	2,08	2,06	1,97	1,92	1,88	1,86	1,84	1,81	1,74	1,69	1,67
28	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,09	2,06	2,04	1,96	1,91	1,87	1,84	1,82	1,79	1,73	1,67	1,65
29	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,08	2,05	2,03	1,94	1,89	1,85	1,83	1,81	1,77	1,71	1,65	1,64
30	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,06	2,04	2,01	1,93	1,88	1,84	1,81	1,79	1,76	1,70	1,64	1,62
40	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,97	1,95	1,92	1,84	1,78	1,74	1,72	1,69	1,66	1,59	1,53	1,51
50	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99	1,95	1,92	1,89	1,87	1,78	1,73	1,69	1,66	1,63	1,60	1,52	1,46	1,44
75	2,73	2,49	2,34	2,22	2,13	2,06	2,01	1,96	1,92	1,88	1,85	1,83	1,80	1,71	1,65	1,61	1,58	1,55	1,52	1,44	1,36	1,34
100	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,89	1,85	1,82	1,79	1,77	1,68	1,62	1,57	1,54	1,52	1,48	1,39	1,31	1,29
200	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,84	1,80	1,77	1,74	1,72	1,62	1,56	1,52	1,48	1,46	1,41	1,32	1,22	1.19
1000	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,80	1,76	1,73	1,70	1,68	1,58	1,52	1,47	1,43	1,41	1,36	1,26	1,13	1,08
∞	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,72	1,69	1,67	1,57	1,51	1,46	1,42	1,40	1,35	1,24	1,11	1

F de Snedecor-Fisher $\alpha = 0{,}01$

g.l. 1 →

g.l. 2 ↓	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500	∞
5	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,82	9,77	9,72	9,55	9,45	9,38	9,33	9,29	9,24	9,13	9,04	9,02
6	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,40	7,30	7,23	7,18	7,14	7,09	6,99	6,90	6,88
7	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,36	6,31	6,16	6,06	5,99	5,94	5,91	5,86	5,75	5,67	5,65
8	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,61	5,56	5,52	5,36	5,26	5,20	5,15	5,12	5,07	4,96	4,88	4,86
9	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,01	4,96	4,81	4,71	4,65	4,60	4,57	4,52	4,41	4,33	4,31
10	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,65	4,60	4,56	4,41	4,31	4,25	4,20	4,17	4,12	4,01	3,93	3,91
11	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,10	4,01	3,94	3,89	3,86	3,81	3,71	3,62	3,60
12	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,86	3,76	3,70	3,65	3,62	3,57	3,47	3,38	3,36
13	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,91	3,86	3,82	3,66	3,57	3,51	3,46	3,43	3,38	3,27	3,19	3,17
14	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,51	3,41	3,35	3,30	3,27	3,22	3,11	3,03	3,00
15	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,37	3,28	3,21	3,17	3,13	3,08	2,98	2,89	2,87
16	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,50	3,45	3,41	3,26	3,16	3,10	3,05	3,02	2,97	2,86	2,78	2,75
17	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,16	3,07	3,00	2,96	2,92	2,87	2,76	2,68	2,65
18	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,32	3,27	3,23	3,08	2,98	2,92	2,87	2,84	2,78	2,68	2,59	2,57
19	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,00	2,91	2,84	2,80	2,76	2,71	2,60	2,51	2,49
20	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,18	3,13	3,09	2,94	2,84	2,78	2,73	2,69	2,64	2,54	2,44	2,42
21	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,12	3,07	3,03	2,88	2,79	2,72	2,67	2,64	2,58	2,48	2,38	2,36
22	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	3,07	3,02	2,98	2,83	2,73	2,67	2,62	2,58	2,53	2,42	2,33	2,31
23	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	3,02	2,97	2,93	2,78	2,69	2,62	2,57	2,54	2,48	2,37	2,28	2,26
24	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,98	2,93	2,89	2,74	2,64	2,58	2,53	2,49	2,44	2,33	2,24	2,21
25	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,94	2,89	2,85	2,70	2,60	2,54	2,49	2,45	2,40	2,29	2,19	2,17
26	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,90	2,86	2,81	2,66	2,57	2,50	2,45	2,42	2,36	2,25	2,16	2,13
27	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99	2,93	2,87	2,82	2,78	2,63	2,54	2,47	2,42	2,38	2,33	2,22	2,12	2,10
28	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,96	2,90	2,84	2,79	2,75	2,60	2,51	2,44	2,39	2,35	2,30	2,19	2,09	2,08
29	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93	2,87	2,81	2,77	2,73	2,57	2,48	2,41	2,36	2,33	2,27	2,16	2,06	2,03
30	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,79	2,74	2,70	2,55	2,45	2,39	2,34	2,30	2,25	2,13	2,03	2,01
40	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,61	2,56	2,52	2,37	2,27	2,20	2,15	2,11	2,06	1,94	1,83	1,80
50	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,63	2,56	2,51	2,46	2,42	2,27	2,17	2,10	2,05	2,01	1,95	1,82	1,71	1,68
75	4,05	3,58	3,27	3,05	2,89	2,76	2,65	2,57	2,49	2,43	2,38	2,33	2,29	2,13	2,03	1,96	1,91	1,87	1,81	1,67	1,55	1,52
100	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,43	2,37	2,31	2,27	2,22	2,07	1,97	1,89	1,84	1,80	1,74	1,60	1,47	1,43
200	3,88	3,41	3,11	2,89	2,73	2,60	2,50	2,41	2,34	2,27	2,22	2,17	2,13	1,97	1,87	1,79	1,74	1,69	1,63	1,48	1,33	1,29
1000	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,27	2,20	2,15	2,10	2,06	1,90	1,79	1,72	1,66	1,61	1,54	1,38	1,19	1,11
∞	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,25	2,18	2,13	2,08	2,04	1,88	1,77	1,70	1,64	1,59	1,52	1,36	1,15	1

F de Snedecor-Fisher $\alpha = 0,001$

g.l. 1 →

g.l. 2 ↓	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500	∞
5	33,2	31,1	29,8	28,8	28,2	27,6	27,2	26,9	26,6	26,4	26,2	26,1	25,9	25,4	25,1	24,9	24,7	24,6	24,4	24,1	23,9	23,79
6	23,7	21,9	20,8	20,0	19,5	19,0	18,7	18,4	18,2	18,0	17,8	17,7	17,6	17,1	16,9	16,7	16,5	16,4	16,3	16,0	15,8	15,75
7	18,8	17,2	16,2	15,5	15,0	14,6	14,3	14,1	13,9	13,7	13,6	13,4	13,3	12,9	12,7	12,5	12,4	12,3	12,2	12,0	11,7	11,70
8	15,8	14,4	13,5	12,9	12,4	12,0	11,8	11,5	11,4	11,2	11,1	10,9	10,8	10,5	10,3	10,1	10,0	9,9	9,8	9,6	9,4	9,33
9	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89	9,72	9,57	9,44	9,33	9,24	8,90	8,69	8,55	8,45	8,37	8,26	8,04	7,86	7,81
10	12,55	11,28	10,48	9,93	9,52	9,20	8,96	8,75	8,59	8,45	8,32	8,22	8,13	7,80	7,60	7,47	7,37	7,30	7,19	6,98	6,81	6,76
11	11,56	10,35	9,58	9,05	8,65	8,35	8,12	7,92	7,76	7,63	7,51	7,41	7,32	7,01	6,81	6,68	6,59	6,52	6,42	6,21	6,04	6,00
12	10,80	9,63	8,89	8,38	8,00	7,71	7,48	7,29	7,14	7,00	6,89	6,79	6,71	6,40	6,22	6,09	6,00	5,93	5,83	5,63	5,46	5,42
13	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80	6,65	6,52	6,41	6,31	6,23	5,93	5,75	5,63	5,54	5,47	5,37	5,17	5,01	4,97
14	9,73	8,62	7,92	7,44	7,08	6,80	6,58	6,40	6,26	6,13	6,02	5,93	5,85	5,56	5,38	5,25	5,17	5,10	5,00	4,81	4,65	4,60
15	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08	5,94	5,81	5,71	5,62	5,54	5,25	5,07	4,95	4,86	4,80	4,70	4,51	4,35	4,31
16	9,01	7,94	7,27	6,80	6,46	6,20	5,98	5,81	5,67	5,55	5,44	5,35	5,27	4,99	4,82	4,70	4,61	4,54	4,45	4,26	4,10	4,06
17	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58	5,44	5,32	5,22	5,13	5,05	4,78	4,60	4,48	4,40	4,33	4,24	4,05	3,89	3,85
18	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39	5,25	5,13	5,03	4,94	4,87	4,59	4,42	4,30	4,22	4,15	4,06	3,87	3,71	3,67
19	8,28	7,27	6,62	6,18	5,85	5,59	5,39	5,22	5,08	4,97	4,87	4,78	4,70	4,43	4,26	4,14	4,06	3,99	3,90	3,71	3,55	3,51
20	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08	4,94	4,82	4,72	4,64	4,56	4,29	4,12	4,00	3,92	3,86	3,77	3,58	3,42	3,38
21	7,94	6,95	6,32	5,88	5,56	5,31	5,11	4,95	4,81	4,70	4,60	4,51	4,44	4,17	4,00	3,88	3,80	3,74	3,64	3,46	3,30	3,26
22	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83	4,70	4,58	4,49	4,40	4,33	4,06	3,89	3,78	3,69	3,63	3,54	3,35	3,19	3,15
23	7,67	6,70	6,08	5,65	5,33	5,09	4,89	4,73	4,60	4,48	4,39	4,30	4,23	3,96	3,79	3,68	3,60	3,53	3,44	3,25	3,10	3,05
24	7,55	6,59	5,98	5,55	5,24	4,99	4,80	4,64	4,51	4,39	4,30	4,21	4,14	3,87	3,71	3,59	3,51	3,45	3,36	3,17	3,01	2,97
25	7,45	6,49	5,89	5,46	5,15	4,91	4,71	4,56	4,42	4,31	4,22	4,13	4,06	3,79	3,63	3,52	3,43	3,37	3,28	3,09	2,93	2,89
26	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48	4,35	4,24	4,14	4,06	3,99	3,72	3,56	3,44	3,36	3,30	3,21	3,02	2,86	2,82
27	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41	4,28	4,17	4,08	3,99	3,92	3,66	3,49	3,38	3,30	3,23	3,14	2,96	2,80	2,75
28	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35	4,22	4,11	4,01	3,93	3,86	3,60	3,43	3,32	3,24	3,18	3,09	2,90	2,74	2,69
29	7,12	6,19	5,59	5,18	4,87	4,64	4,45	4,29	4,16	4,05	3,96	3,88	3,80	3,54	3,38	3,27	3,18	3,12	3,03	2,84	2,68	2,64
30	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24	4,11	4,00	3,91	3,82	3,75	3,49	3,33	3,22	3,13	3,07	2,98	2,79	2,63	2,59
40	6,59	5,70	5,13	4,73	4,44	4,21	4,02	3,87	3,75	3,64	3,55	3,47	3,40	3,15	2,98	2,87	2,79	2,73	2,64	2,44	2,28	2,23
50	6,34	5,46	4,90	4,51	4,22	4,00	3,82	3,67	3,55	3,44	3,35	3,27	3,20	2,95	2,79	2,68	2,60	2,53	2,44	2,25	2,07	2,03
75	6,01	5,16	4,62	4,24	3,96	3,74	3,56	3,42	3,30	3,19	3,10	3,03	2,96	2,71	2,55	2,44	2,35	2,29	2,19	1,99	1,81	1,75
100	5,86	5,02	4,48	4,11	3,83	3,61	3,44	3,30	3,18	3,07	2,99	2,91	2,84	2,59	2,43	2,32	2,24	2,17	2,08	1,87	1,67	1,62
200	5,63	4,81	4,29	3,92	3,65	3,43	3,26	3,12	3,00	2,90	2,82	2,74	2,67	2,42	2,26	2,15	2,07	2,00	1,90	1,68	1,46	1,39
1000	5,46	4,65	4,14	3,78	3,51	3,30	3,13	2,99	2,87	2,77	2,69	2,61	2,54	2,30	2,14	2,02	1,94	1,87	1,77	1,53	1,27	1,15
∞	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96	2,84	2,74	2,66	2,58	2,51	2,27	2,10	1,99	1,90	1,84	1,73	1,49	1,21	1

Tema 11 : Planificación de estudios estadísticos. Clases de estudios.

Los descubrimientos o avances científicos pueden ser fruto de
1) la casualidad, muy a menudo unida a una intuición genial. Por ejemplo, el descubrimiento de los Rx, la penicilina, el yodo, la ley de la gravedad....
2) la búsqueda de soluciones a problemas, como la necesidad de nuevos medicamentos o nuevos combustibles.
3) la curiosidad teórica, con Einstein como uno de los mejores ejemplos.

El primer camino es excepcional, no porque no se den ocasiones, sino porque la mayoría de las personas no reconocen la trascendencia de la observación. La suerte sólo favorece a los preparados (Pasteur). Los otros dos caminos son los habituales y requieren un estudio planificado.

Etapas fundamentales de un estudio

En un estudio planificado se pueden distinguir 5 etapas fundamentales: 1 planteamiento, 2 información, 3 formulación de la hipótesis, 4 realización u obtención de datos y 5 análisis de resultados y conclusiones.
Esta distinción se hace a efectos teóricos y didácticos, pues en la práctica al comienzo del trabajo se imbrican las tres primeras etapas y sólo al cabo de un tiempo quedan claramente definidas, cosa que inexcusablemente debe de ocurrir antes de iniciar el paso 4º, la realización. Veamos estas etapas con más detalle:

1) **PLANTEAMIENTO** : qué se va a estudiar, por qué, para qué, cómo, etc
   El "cómo" incluye
   a) el diseño de la investigación: lo que habitualmente se conoce en los trabajos científicos como material y métodos, p.e. el nº de individuos a estudiar, las características que deben reunir, el procedimiento de elección, tratamiento aplicado, variables a medir, etc
   b) las necesidades de material, personal y dinero.

   Como ya se ha dicho el planteamiento inicial es provisional, pudiendo ser modificado en función de los pasos 2 y 3.
2) **INFORMACION** : es preciso saber lo máximo posible sobre el tema de la investigación, consultando libros y revistas especializadas. Es lo que se llama "revisión bibliográfica" o "revisión de la literatura".
   Este material debe ser valorado críticamente. Ante cada trabajo concreto hay que hacerse una serie de preguntas. ¿quien lo ha escrito? , ¿donde? , ¿cuando?, ¿el material y el método utilizados son correctos?, ¿están justificadas las conclusiones?, etc... El motivo de esta valoración crítica es que es muy, muy difícil hacer bien un trabajo científico, por lo que la inmensa mayoría tienen errores y deficiencias más o menos transcendentes.
   Tras este examen habrá cosas claras y generalmente aceptadas, mientras que otras serán inciertas, dudosas o controvertidas. Se tomará buena nota de los fallos observados en otros investigadores para no incurrir en ellos.
3) **HIPOTESIS** : es la explicación provisional de unos hechos. Al concluir la investigación se verá si es o no cierta ("verificación" de la hipótesis). Los estudios puramente descriptivos no tienen hipótesis, aunque pueden servir de base para formular hipótesis.
4) **REALIZACION U OBTENCION DE DATOS (RECOGIDA DE LA INFORMACION)**
   Para ello se va cumpliendo exactamente lo previsto en el punto "Material y métodos" del paso nº 1. Una vez recogidos todos los datos se clasifican y ordenan siguiendo las normas de la Estadística Descriptiva. Es importante buscar posibles errores de ejecución y desechar todo lo que no se ajuste exactamente al método previsto.
5) **ANALISIS DE LOS RESULTADOS Y CONCLUSIONES**
   Se aplica el método de análisis estadístico que corresponda al tipo de datos y al objetivo de la investigación. Así se verifica la hipótesis de trabajo, es decir se confirma o se desecha. Las

hipótesis no confirmadas también tiene su valor. Así, puede concluirse que un nuevo medicamento no es más eficaz que los que había, que una nueva técnica no mejora la actual, etc. Todo ello permitirá sacar CONCLUSIONES. Hay que distinguir entre las conclusiones estadísticas, que como se verá en su momento llevan anejo un juicio de significación y si es posible un juicio de causalidad, y las conclusiones del estudio que se basan en las anteriores. Es conveniente recordar que las conclusiones estadísticas lo son a nivel de grupo, no a nivel individual. Son válidas para la inmensa mayoría de los individuos, no para todos. "La estadística no es una ciencia exacta".

Un error frecuente es sacar conclusiones basadas en la información previa, no en el estudio

Clases de estudios estadísticos

Se pueden clasificar desde distintos puntos de vista:

- **en función del nº de variables:**
  - E. de **INFORMACION**: estudio de una variable
    - DESCRIPTIVOS: tabulación, representación gráfica, índices estadísticos...
    - de ESTIMACION: estimar parámetros de una población a partir de una muestra
    - de CONFORMIDAD: valorar si una muestra puede proceder de una población determinada
  - E. de **INVESTIGACION** O COMPARATIVOS: diferencias o relaciones entre dos o más variables
    - EXPERIMENTALES
      - Clásicos: 1 variable controlada y el resto aleatorias
      - Factoriales: 2 ó más variables controladas y el resto aleatorias
    - de OBSERVACION: todas las variables son aleatorias.

Sólo los estudios experimentales permiten una interpretación causal

- **en función del momento en que se generan los datos:**
  - Estudios <u>RETROSPECTIVOS</u> o históricos. Los datos ya se han generado cuando se planifica, por lo que los métodos previstos en "material y métodos" pueden no haber sido observados exactamente. p.e. se revisan las historias clínicas de 1000 pacientes que tomaron el medicamento M para ver los efectos secundarios que presentaron.

    A este grupo pertenecen los estudios caso-control: un grupo de individuos afectados se compara con otro u otros no afectados para investigar el nivel de exposición a determinados factores que podrían ser causales o protectores. Cada caso se empareja con uno o más controles, que por lo demás deben ser lo más parecidos posible a los casos (sexo, edad, etc). Es la herramienta de trabajo clásica de los estudios epidemiológicos, p.e., en el caso de una intoxicación alimenticia en una boda. Su parámetro típico es la razón de probabilidad u ODDS RATIO (OR), que veremos en otro tema.
  - Estudios <u>PROSPECTIVOS</u> o de futuro. Los datos se generan después de la planificación del estudio y como consecuencia del mismo. p.e. a partir de hoy se van a recoger los efectos secundarios en mil pacientes consecutivos que toman el medicamento M.

    A este grupo pertenecen los estudios de cohortes, típicos de estudios epidemiológicos, mucho menos usados que los de caso control. Son difíciles y caros y llevan más tiempo. Se seleccionan individuos expuestos y no expuestos a un factor y a lo largo del tiempo se ve si enferman o no. Su parámetro típico es el cociente de riesgo o riesgo relativo (RR), que veremos también en otro tema.

- **en función de los individuos:**
  - Estudios con datos independientes. Los individuos están repartidos en dos o más grupos o muestras; cada individuo sólo forma parte de un grupo. p.e. se prueba el medicamento A en 100 individuos y el B en otros 100.
  - Estudios con datos apareados. todos los individuos forman parte de todos los grupos. El orden por el que entran en cada uno de los grupos se determina al azar. p.e. 100 pacientes reciben en momento dado el medicamento A y en otro momento el B y se comparan sus efectos. Los 100 pacientes forman parte del grupo medicamento A y también del grupo medicamento B.
- **en función del conocimiento de los detalles y resultado del estudio:**
  - Abiertos. Los que realizan el estudio, los que lo valoran y, si son conscientes, también los individuos conocen los grupos y el tratamiento que reciben.
  - Ciegos. Quien valora los resultados desconoce a que grupo pertenecen los individuos y por tanto el tratamiento recibido.
  - Doble ciegos. Ese desconocimiento se extiende a los que realizan el estudio, a los que lo valoran y a los individuos, si son conscientes. Sólo el director del estudio, que no hace la valoración, revela al final todos los detalles.
- **en función del lugar en que se realiza el estudio:**
  - unicéntricos : todo el estudio se realiza por el mismo equipo investigador
  - multicéntricos: el estudio se realiza simultáneamente en diversos sitios por diversos investigadores siguiendo un diseño común.
- **en función del método experimental:**
  - con tratamiento activo. Se da el producto que se investiga.
  - con placebo. Se aplica un tratamiento inactivo, sin efecto, con el mismo aspecto externo que el tratamiento activo. Esto se aplica sólo a humanos y lógicamente el individuo no sabe lo que está tomando.

En los últimos años las revistas científicas más prestigiosas han introducido de forma obligatoria la “Declaración de intereses”: los autores declaran si tienen o han tenido alguna relación laboral, comercial, de asesoría o de mecenazgo con personas, empresas o instituciones que tengan algo que ver con el estudio. Es decir, si hay o no hay “conflicto de intereses”.

Los mejores estudios son los unicéntricos, experimentales, prospectivos, doble ciegos, incluyendo placebo y si es posible con datos apareados.

Tema 12 : Recogida de la información, Técnicas de muestreo. Errores de los muestreos.

---Una muestra debe ser representativa

Ya vimos en el tema 1 que las muestras deben ser representativas de la población de la que proceden y que la mejor garantía de conseguirlo es un tamaño adecuado de la muestra y la elección al azar de los individuos, es decir, **una muestra aleatoria de tamaño adecuado**. Es un punto crucial.
Esta representatividad puede verse afectada, además de por un tamaño insuficiente, por los llamados **factores de sesgo**, como deficiencias de la aleatoriedad (¿tienen realmente todos los individuos la misma probabilidad de salir elegidos?), errores muestrales extremos y errores personales e instrumentales.

---Origen de la muestra

La población de la que procede la muestra es la **población muestreada o población origen**, que idealmente debe coincidir con **la población objetivo** del estudio, lo que no siempre ocurre por la existencia de **factores selectivos** más o menos intensos. Es posible que el investigador no se de cuenta de esta situación y pueda llegar, honestamente, a conclusiones erróneas.



Siempre hay que comprobar que la población muestreada es realmente la población objetivo

Ejemplo: en los años 50 se realizó en Barcelona un estudio epidemiológico muy importante sobre tuberculosis, que estaba entonces muy extendida. Los datos se obtuvieron de una muestra tomada del Dispensario Antituberculoso. Los resultados se presentaron como reflejo del estado de la tuberculosis en la ciudad de Barcelona. Pronto surgieron críticas al estudio. ¿La muestra era realmente representativa de los tuberculosos catalanes?. ¿O sólo de los pobres?. Los más pudientes y algunos más pobres que hicieron un esfuerzo económico eran atendidos en consultas y clínicas privadas. Y era de sobra sabido la influencia del estado social en la evolución de esta enfermedad. Muy probablemente la muestra estaba contaminada por un factor selectivo: la situación económica.

---Tamaño de la muestra

Depende fundamentalmente de 4 factores: 1) tamaño de la población, 2) dispersión o variabilidad de los individuos de la población, 3) margen de error que estemos dispuestos a admitir y 4) nivel de significación o confianza elegidos.

Para calcular el tamaño muestral se dispone de fórmulas, que nos orientan sobre el mismo. Siempre se cogen más individuos de los calculados, para compensar posibles fallos. También se dispone de tablas, sobre todo para estimaciones de porcentajes, que no veremos. En la práctica a partir de un tamaño poblacional de 10.000 se pueden usar las fórmulas de "población infinita", que son más sencillas. Dicho de otra forma: a efectos prácticos una población se puede considerar como infinita a partir de un tamaño de 10.000 (hay autores que elevan este tamaño a 60.000).
En las fórmulas aparece c^2. Es el valor de c de la DN tipificada que corresponde al nivel de significación elegido. El nivel de significación, cuyo símbolo es α, expresa el riesgo estadístico de error, el llamado "error tipo 1". Por consenso se consideran significativos los valores de α de 0'05 para abajo. Los programas estadísticos de ordenador calculan este riesgo exactamente. Para cálculos manuales se toman tradicionalmente tres puntos de referencia para α : 0'05 (ó 5%) , 0'01 (ò 1%) y 0'001 (ó 1%º)., que se corresponden con valores de c de 1'96 , 2'53 y 3'30 respectivamente. Si no se exige o desea otro nivel, se toma de oficio el de 0'05 y por tanto c = 1'96.

---Fórmulas

1) para una estimación

	Población finita	Población infinita
media	$N = \dfrac{c^2 * Np * s^2}{Np * k^2 + c^2 * s^2}$	$N = \left(\dfrac{c * s}{k}\right)^2$
p ó %	$N = \dfrac{c^2 * Np * p * q}{(Np-1) * k^2 + c^2 * p * q}$	$N = \dfrac{c^2 pq}{k^2}$

2) para contraste de variables (N por muestra)
- de medias : $N = 13 * s^2 / d^2$
- de 2 proporciones o porcentajes : $N = 6'5(p_1q_1+p_2q_2)/d^2$

N es el tamaño muestral, N_p el tamaño de la población, k el error máximo admitido, s^2 la varianza de la población, real o estimada a partir de un estudio piloto o incluso de una forma más simple por la fórmula $s^2 \approx (R/4)^2$, siendo R el Recorrido. La "c" es el valor de referencia de la DN tipificada correspondiente al nivel de significación elegido. La "d" es la diferencia mínima que queremos probar entre los porcentajes o medias contrastadas.
En el caso de estimaciones p y q toman su valor real en la población si se conoce; si no, se les da el valor más desfavorable y que conduce a un tamaño mayor: 0'5 a cada una. En el caso de contraste de muestras se procede de la misma forma: dar a cada p y q su valor real, si es conocido y si no, darles el valor de 0'5.
Si los datos son apareados o se trata de una prueba de conformidad, N se divide por 2.

---Recogida de los datos

Los datos se recogen por
1) observación, directa o con aparatos.
2) interrogatorio, que puede ser directo (entrevista) o indirecto (cuestionario). Es típico de encuestas. Presupone preguntas neutrales y por parte del interrogado buena memoria y buena fe.

---Métodos de obtención de muestras al azar

Hay diversos tipos de muestras aleatorias:

1. **Muestras de azar simple o aleatoria elemental**. Presupone lista de todos los individuos, numerados. La unidad muestral es el individuo. Los individuos se eligen por sorteo o utilizando una tabla de números al azar (ver una muy sencilla al final del tema).
2. **Muestras sistemáticas**. Es una variante de la anterior con un procedimiento de elección simplificado. Hay que calcular el coeficiente de elevación (Tamaño de la población dividido por el tamaño de la muestra). Luego se elige al azar un número menor que dicho coeficiente, que será el primer individuo de la muestra. A ese número se la sumando el coeficiente de elevación y así nos va dando los individuos hasta alcanzar el tamaño previsto de la muestra. Por ejemplo: tamaño de la población 1000; tamaño de la muestra 100 ; coeficiente de elevación 1000/100 = 10. Se elige al azar un número menor de 10 y sale el 6. La muestra la compondrán los individuos de la lista cuyos números sean el 6 , 16 , 26 , 36 , 46 , ... hasta el 996.
3. **Muestras estratificadas**. Se hacen estratos de la población, que son grupos homogéneos de individuos, con poca variación intragrupo. Por ejemplo, hombres y mujeres, grupos de edad, grupos raciales, regiones de un país, factores de riesgo. etc. Fijados los estratos se eligen de forma proporcional y al azar los individuos que formarán la muestra. Aquí también la unidad muestral es el individuo y se necesita un listado de la población. son muy utilizadas en investigaciones clínicas.
4. **Muestras de conglomerados.** Los conglomerados son grupos naturales y heterogéneos de individuos. De entrada no se conocen los individuos, sino los conglomerados, que son la unidad muestral. Por ejemplo, tenemos una lista de escuelas o de hospitales (que son los conglomerados) ; se eligen al azar los que hagan falta y una vez en ellos se eligen al azar los individuos necesarios.
5. **Muestras combinadas**. Es una mezcla de estratos y conglomerados.

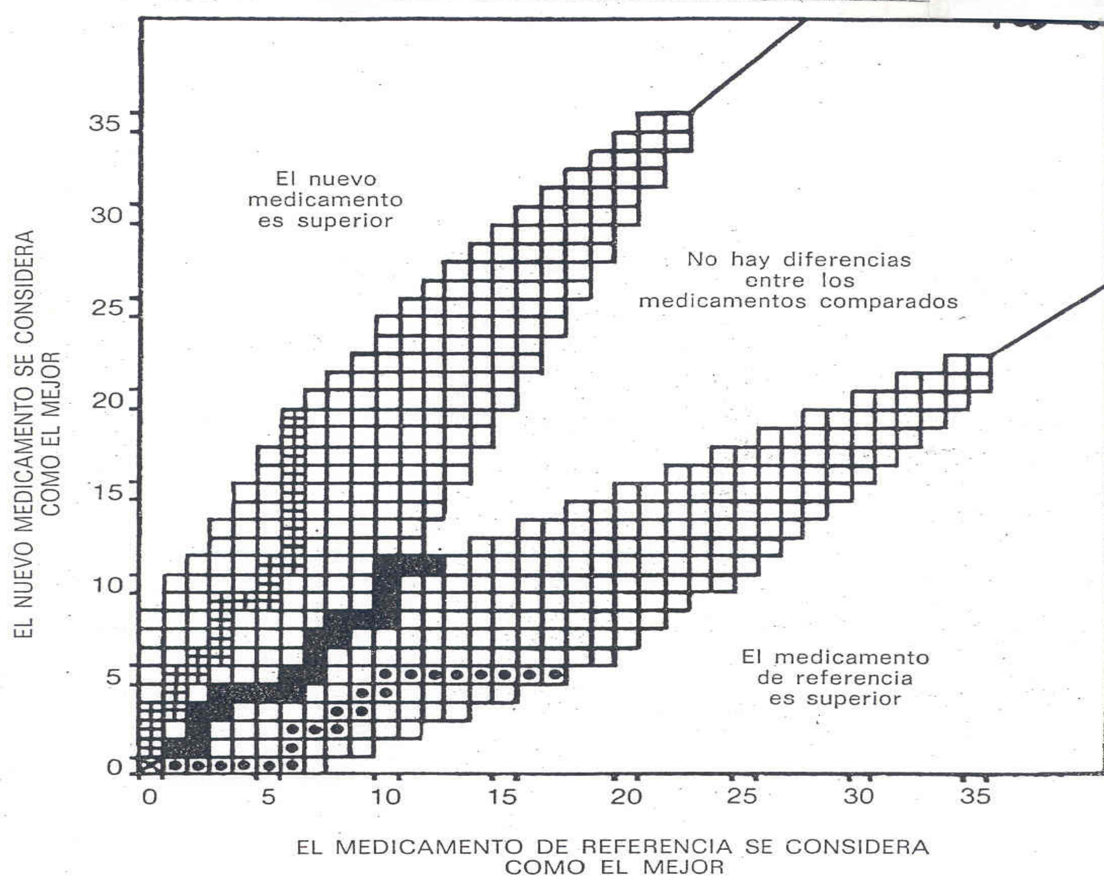
<u>Ejemplos</u>: Deseamos estudiar el nivel de plomo en la sangre de los niños de 3º de ESO en la región R. Sabemos que son 4000 niños, que acuden a 200 escuelas y cada clase tiene 20 alumnos. Tenemos un listado de los 4000 alumnos y un listado de las escuelas. 40 escuelas están en poblaciones grandes, 120 en medianas y 80 en pequeñas- Supongamos que necesitamos una muestra de tamaño 400. ¿Cómo obtenerla?

1. *Muestra al azar*. De la lista de los 4000 niños se sacan al azar (sorteo o por la tabla de números al azar) los 400 que se necesitan.
2. *Muestra sistemática*. Necesitamos también la lista de los 4000 alumnos. Coeficiente de elevación : 4000/400=10. Se elige al azar un número <10 y sale el 3. Por tanto saldrán elegidos para formar parte de la muestra los alumnos con los números 3, 13, 23, 33, 43,......y así hasta el 3993.
3. *Muestra estratificada*. Hay indicios de que el tamaño de las ciudades y pueblos puede ser de importancia en el estudio. Elegimos 3 estratos representativos y les asignamos un porcentaje (fruto del estudio de la situación): ciudades o pueblos grandes, de los que sacaremos el 20% de la muestra; medianos con el 60% y pequeños con el 20%. Esto equivale a tomar 80 alumnos del estrato grande, 240 del mediano y 80 del pequeño. Su elección se hace por el método 1 ó el 2.
4. *Muestra de conglomerados*. Aquí no hay lista de alumnos, sólo de escuelas. Se eligen al azar 20 escuelas y se toman los 20 alumnos de cada una de ellas.
5. *Muestra combinada*. Une 3 y 4. Agrupamos las escuelas (que son los conglomerados) por estratos de tamaño poblacional (40, 120, 40) y se eligen el 10% de cada estrato, o sea 20 , 12 y 4 escuelas respectivamente. tomando los 20 alumnos de cada una de estas escuelas tenemos los 400 necesarios.

---Otras formas de obtener muestras

En investigaciones clínicas se utiliza con frecuencia la llamada **asignación al azar**, que evita elecciones subjetivas. Por ejemplo, en estudios en que cada paciente nuevo debe ser asignado a un grupo de tratamiento distinto; se dispone de una serie de sobres cerrados en los que está el tratamiento a recibir y cuando llega el paciente se coge un sobre y se le aplica el tratamiento que indica.

En el **análisis secuencial** no es necesario siquiera conocer previamente el tamaño muestral. Los datos se comparan por parejas, uno del grupo que podemos llamar A y otro del grupo B. Hay 3 resultados posibles: A es mejor, B es mejor y ninguno es mejor (0). Se utiliza una gráfica en V, como la que sigue, que sirve para α =0,05. Se van rellenado casillas con los datos que vamos obteniendo. Se empieza por el vértice de la V. Si A es mejor se rellena la casilla superior, si es mejor B la casilla de la derecha y si no hay diferencias no se rellena ninguna casilla. Llega un momento en que nos salimos del gráfico por algún sitio. Por arriba si A es mejor, por abajo si B es mejor y por el centro si no hay diferencias.



Supongamos que queremos ver si un nuevo medicamento (A) es superior al que actualmente se utiliza (B) en el tratamiento de la migraña. Cada paciente recibe en un orden prefijado al azar un medicamento, en una ocasión A y B en otra. Luego informa de cual ha sido más eficaz. Se obtiene lo siguiente:

paciente: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 ..
mejor: A A B B A A 0 0 A A A B A A 0 B A A A A B A B A 0 0..

paciente: ... 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
mejor: .... B A A A B 0 0 B A A 0 B A A A 0 A 0 A

En el paciente 45 nos salimos de la V por arriba. Por tanto A es mejor.

---Errores de los muestreos

I. PROPIOS DE LA MUESTRA
   i. muestra no representativa
   ii. *ERROR MUESTRAL*, que es inevitable y se debe a la variabilidad natural. Se puede medir hasta donde puede llegar. Lo veremos enseguida.

II. EXTRAÑOS A LA MUESTRA
   i. personales (del observador), que dependen de su preparación, estado psico-físico, ambiente, etc. Hay variaciones intraobservador e interobservador.
   ii. sistemáticos (del método de medida). Dependen de su sensibilidad, precisión y exactitud.

Sesgos de recuerdo ("recall bias")

Los pacientes son reiteradamente preguntados por la existencia de factores de riesgo y los suelen recordar muy bien. Cosa que no ocurre con los controles en un estudio caso-control.

---Disminución de los errores

--los del observador, mediante una buena preparación, condiciones adecuadas de trabajo y utilización de controles de calidad.
--los del método, mediante aparatos de calidad, buen mantenimiento, controles de calidad, buenos cuestionarios.

---ERROR MUESTRAL (E)

Si sacamos de una población diversas muestras y calculamos uno o más parámetros, veremos que no obtenemos exactamente los mismos resultados. Esto se debe a la existencia de un error, el error muestral, que es inevitable, pero que puede ser valorado, ya que los parámetros obtenidos de muestras repetidas de una misma población (>30) siguen la ley normal aunque la población de origen no sea normal. Y por tanto tienen su margen de variación, cuyo máximo puede ser medido. Es el error muestral.

E = c*e ó t*e, siendo e el llamado error estándar. Si la muestra es <30 se utiliza t, la t de Student, y si es grande (≥30) la c de la DN.

---ERROR ESTANDAR (e)

Es la desviación estándar de la distribución de los parámetros estadísticos muestrales (media, %, etc.) cuando se extraen repetidas muestras. No se debe confundir con la desviación estándar de una muestra (s). Se han encontrado fórmulas con las que a partir de una sola muestra se puede calcular ya el error estándar:

para una media: $$e = \frac{s}{\sqrt{N}}$$

para un porcentaje: $$e = \sqrt{\frac{pq}{N}}$$

TABLA VII

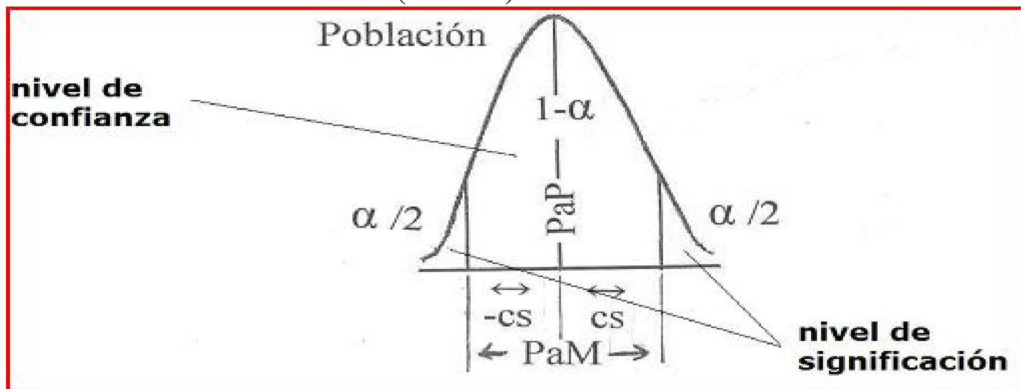
TABLA DE NUMEROS ALEATORIOS

10 09 73 25 33	76 52 01 35 86	34 67 35 48 76	80 95 90 91 17	39 29 27 49 45
37 54 20 48 05	64 89 47 42 96	24 80 52 40 37	20 63 61 04 02	00 82 29 16 65
08 42 26 89 53	19 64 50 93 03	23 20 90 25 60	15 95 33 47 64	35 08 03 36 06
99 01 90 25 29	09 37 67 07 15	38 31 13 11 65	88 67 67 43 97	04 43 62 76 59
12 80 79 99 70	80 15 73 61 47	64 03 23 66 53	98 95 11 68 77	12 17 17 68 33
66 06 57 47 17	34 07 27 68 50	36 69 73 61 70	65 81 33 98 85	11 19 92 91 70
31 06 01 08 05	45 57 18 24 06	35 30 34 26 14	86 79 90 74 39	23 40 30 97 32
85 26 97 76 02	02 05 16 56 92	68 66 57 48 18	73 05 38 52 47	18 62 38 85 79
63 57 33 21 35	05 32 54 70 48	90 55 35 75 48	28 46 82 87 09	83 49 12 56 24
73 79 64 57 53	03 52 96 47 78	35 80 83 42 82	60 93 52 03 44	35 27 38 84 35
98 52 01 77 67	14 90 56 86 07	22 10 94 05 58	60 97 09 34 33	50 50 07 39 98
11 80 50 54 31	39 80 82 77 32	50 72 56 82 48	29 40 52 42 01	52 77 56 78 51
83 45 29 96 34	06 28 89 80 83	13 74 67 00 78	18 47 54 06 10	68 71 17 78 17
88 68 54 02 00	86 50 75 84 01	36 76 66 79 51	90 36 47 64 93	29 60 91 10 62
99 59 46 73 48	87 51 76 49 69	91 82 60 89 28	93 78 56 13 68	23 47 83 41 13
65 48 11 76 74	17 46 85 09 50	58 04 77 69 74	73 03 95 71 86	40 21 81 65 44
80 12 43 56 35	17 72 70 80 15	45 31 82 23 74	21 11 57 82 53	14 38 55 37 63
74 35 09 98 17	77 40 27 72 14	43 23 60 02 10	45 52 16 42 37	96 28 60 26 55
69 91 62 68 03	66 25 22 91 48	36 93 68 72 03	76 62 11 39 90	94 40 05 64 18
09 89 32 05 05	14 22 56 85 14	46 42 75 67 88	96 29 77 88 22	54 38 21 45 98
91 49 91 45 23	68 47 92 76 86	46 16 28 35 54	94 75 08 99 23	37 08 92 00 48
80 33 69 45 98	26 94 03 68 58	70 29 73 41 35	53 14 03 33 40	42 05 08 23 41
44 10 48 19 49	85 15 74 79 54	32 97 92 65 75	57 60 04 08 81	22 22 20 64 13
12 55 07 37 42	11 10 00 20 40	12 86 07 46 97	96 64 48 94 39	28 70 72 58 15
63 60 64 93 29	16 50 53 44 84	40 21 95 25 63	43 65 17 70 82	07 20 73 17 90
61 19 69 04 46	26 45 74 77 74	51 92 43 37 29	65 39 45 95 93	42 58 26 05 27
15 47 44 52 66	95 27 07 99 53	59 36 78 38 48	82 39 61 01 18	33 21 15 94 66
94 55 72 85 73	67 89 75 43 87	54 62 24 44 31	91 19 04 25 92	92 92 74 59 73
42 48 11 62 13	97 34 40 87 21	16 86 84 87 67	03 07 11 20 59	25 70 14 66 70
23 52 37 83 17	73 20 88 98 37	68 93 59 14 16	26 25 22 96 63	05 52 28 25 62
04 49 35 24 94	75 24 63 38 24	45 86 25 10 25	61 96 27 93 35	65 33 71 24 72
00 54 99 76 54	64 05 18 81 59	96 11 96 38 96	54 69 28 23 91	23 28 72 95 29
35 96 31 53 07	26 89 80 93 54	33 35 13 54 62	77 97 45 00 24	90 10 33 93 33
59 80 80 83 91	45 42 72 68 42	83 60 94 97 00	13 02 12 48 92	78 56 52 01 06
46 05 88 52 36	01 39 09 22 86	77 28 14 40 77	93 91 08 36 47	70 61 74 29 41
32 17 90 05 97	87 37 92 52 41	05 56 70 70 07	86 74 31 71 57	85 39 41 18 38
69 23 46 14 06	20 11 74 52 04	15 95 66 00 00	18 74 39 24 23	97 11 89 63 38
19 56 54 14 30	01 75 87 53 79	40 41 92 15 85	66 67 43 68 06	84 96 28 52 07
45 15 51 49 38	19 47 60 72 46	43 66 79 45 43	59 04 79 00 33	20 82 66 95 41
94 86 43 19 94	36 16 81 08 51	34 88 88 15 53	01 54 03 54 56	05 01 45 11 76
98 08 62 48 26	45 24 02 84 04	44 99 90 88 96	39 09 47 34 07	35 44 13 18 80
33 18 51 62 32	41 94 15 09 49	89 43 54 85 81	88 69 54 19 94	37 54 87 30 43
80 95 10 04 06	96 38 27 07 74	20 15 12 33 87	25 01 62 52 98	94 62 46 11 71
79 75 24 91 40	71 96 12 82 96	69 86 10 25 91	74 85 22 05 39	00 38 75 95 79
18 63 33 25 37	98 14 50 65 71	31 01 02 46 74	05 45 56 14 27	77 93 89 19 36
74 02 94 39 02	77 55 73 22 70	97 79 01 71 19	52 52 75 80 21	80 81 45 17 48
54 17 84 56 11	80 99 33 71 43	05 33 51 29 69	56 12 71 92 55	36 04 09 03 24
11 66 44 98 83	52 07 98 48 27	59 38 17 15 39	09 97 33 34 40	88 46 12 33 56
48 32 47 79 28	31 24 96 47 10	02 29 53 68 70	32 30 75 75 46	15 02 00 99 94
69 07 49 41 38	87 63 79 19 76	35 58 40 44 01	10 51 82 16 15	01 84 87 69 38

Tema 13 : Intervalos de probabilidad y confianza. Hipótesis y decisiones estadísticas.

---Intervalo de probabilidad (IP)

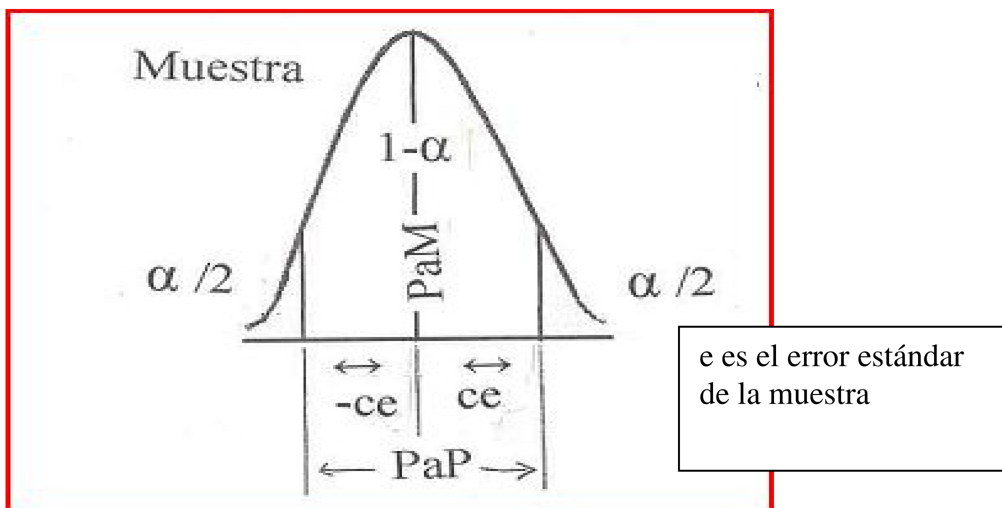
Permite predecir el comportamiento de las muestras.
Si de una población se sacan infinitas muestras y se calcula en ellas un parámetro (media, %, etc.), los resultados varían siguiendo una DN y la media de todos ellos coincide con el parámetro de la población (PaP o PP).
La probabilidad de que el parámetro de una muestra (PaM o PM) esté dentro de un determinado intervalo de valores es 1-α y la probabalidad de estar fuera de ese intervalo es α. A 1-α se le llama **nivel de confianza** y a α **nivel de significación**. La suma de ambos niveles vale 1 (ó 100%).



α la fijamos nosotros y habitualmente se manejan tres puntos de referencia: 0,05 (ó 5%), 0,01(ó 1%) y 0,001 (ó 1‰) Por tanto los correspondientes puntos de referencia del nivel de confianza son: 0,95 (95%) ; 0,99 (99%) ; 0,999 (99,9%).
A esos tres valores de α le corresponden en la DN los siguientes valores de c: 1,96 ; 2,58 y 3,30 , respectivamente

--Intervalo de confianza (IC)

Se obtiene a partir de una muestra en la que calculamos un parámetro y , aplicando la fórmula correspondiente, también un intervalo, en el que estará el verdadero valor del parámetro en la población al nivel de confianza que se elija.



e es el error estándar de la muestra

---Las **PRUEBAS DE HIPOTESIS**, típicas de la Estadística Inferencial se dividen en cuatro grandes clases:

1. **Pruebas de estimación**. A partir del parámetro de la muestra hacemos una estimación de ese parámetro en la población calculando el intervalo de confianza.
2. **Pruebas de conformidad**, que permiten verificar si el parámetro calculado en una muestra puede proceder de una población determinada. Puede proceder si ese parámetro está dentro del intervalo de probabilidad de la población. Estas pruebas contestan a las preguntas: ¿Puede proceder...de...?, ¿Es conforme...con...?

Pruebas de contraste de variables:

3**.--Pruebas de relación o dependencia**. Permiten verificar si dos o más variables están relacionadas o son independientes. Contestan a las preguntas: ¿Hay una relación entre las variables? , ¿los valores de Y dependen de los de X?,...

4.--**Pruebas de comparación**, que permiten saber si las diferencias observadas entre dos o más muestras se deben al azar, en cuyo caso no existen diferencias de importancia estadística; son muestras de la misma población y están dentro de su intervalo de probabilidad. Contestan a la pregunta: Los datos de las muestras que comparamos son más o menos iguales o difieren significativamente?

Cuando los datos son independientes, relación y comparación son lo mismo, simples variantes de enfoque del mismo problema, y se resuelven utilizando las mismas fórmulas. En cambio, si los datos son apareados, las dos pruebas son esencialmente distintas y se resuelven con fórmulas distintas. ¡Hay que hacerse las preguntas correspondientes para elegir el camino adecuado!.

---Metódica de las pruebas de hipótesis

1. Se formula la hipótesis estadística
2. Se aplica la prueba o test estadístico que corresponda
3. En función de los resultados se toma una decisión estadística.

*** * *** La **HIPOTESIS ESTADISTICA** inicial es la **Hipótesis nula (H_0) de igualdad o no relación** entre las variables contrastadas. Dice que las diferencias de los parámetros de las variables no son diferencias importantes, que son debidas a las fluctuaciones del azar. O que no hay relación entre ellos. Todos proceden de la misma población, están dentro de su intervalo de probabilidad, también llamado zona de no rechazo de H_0. Ya sabemos que un valor cualquiera tiene una probabilidad 1-α (el nivel de confianza) de estar en esa zona.

Si el resultado de la prueba, y sólo entonces, conduce al rechazo de H_0 , aparece y se acepta la **Hipótesis alternativa (H_1) de no igualdad o relación** entre las variables contrastadas. Las diferencias observadas no se explican por el azar, las muestras proceden de poblaciones distintas, ya que quedan fuera del IP, en la llamada zona de rechazo de H_0 , cuya p es el nivel de significación α.

No hay que confundir la hipótesis del trabajo con la hipótesis estadística. Supongamos que hacemos un estudio esperando que un nuevo método terapéutico sea superior al clásico. Esta será la hipótesis del estudio. La hipótesis estadística será H_0 , o sea, que no hay diferencias de importancia estadística entre ambos métodos. Si la prueba estadística conduce al rechazo de H_0 , entonces se acepta H_1, que dirá que sí que hay diferencias significativas.

H_1 es habitualmente doble (pruebas bilaterales): las diferencias pueden estar a un lado u otro ; la relación puede ser positiva o negativa**. ¡Siempre que se acepte H_1 hay que indicar el sentido!**. En ocasiones, poco frecuentes en la práctica, puede interesar sólo uno de los sentidos (pruebas unilaterales).

*** * *** Las **pruebas estadísticas** se irán viendo en temas sucesivos.

** * La **decisión estadística** se toma en general siguiendo estos pasos:

1) **se aplica la prueba estadística** correspondiente, obteniendo un resultado, que para unificar el lenguaje llamaremos **z** , nombre arbitrario (podría llamarse de cualquier otra forma) que evita las confusiones que origina el hábito muy extendido de llamar a los resultados de las pruebas con el nombre de la distribución de referencia con que se valora los resultados (t de Student, χ^2, etc.). La prueba estadística se elige en función de la variable (CL o CT), de la naturaleza de los datos (independientes o apareados), del tamaño de la muestra y del cumplimiento de determinadas condiciones de aplicación.
2) **se busca el valor de referencia** (c de la DN, t , , χ^2 , F...) correspondiente al nivel de significación propuesto o en su defecto a 0,05.
3) **se compara z(en valor absoluto, con el valor de referencia (Ref.)** :
   a. **si $|z| <$ Ref. : no se puede rechazar H_0**. No se han encontrado diferencias estadísticamente significativas entre los grupos contrastados o no hay relación entre ellos, son independientes. Realmente es más correcto decir que no se puede rechazar H_0, que decir, cosa que se hace con frecuencia, que se acepta H_0 o que H_0 es verdadera. Nos quedamos con ella porque no podemos rechazarla. Es como una absolución por falta de pruebas. Se indica por n.s. (no significativo) ó $p>0,05$.
   b. si $|z| \geq$ Ref. : se rechaza H_0 y se acepta H_1 a ese nivel de significación. Hay diferencias o una relación con significación estadística. El sentido de las diferencias o de la relación, que siempre se debe dar, se deduce de los datos y parámetros. Se simboliza por $p< \alpha$ (el que corresponda).

En las pruebas de estimación y de conformidad, si no se dice otra cosa, sólo se toma el nivel de significación de 0,05. En las pruebas de contraste, si se supera un nivel hay que probar con el siguiente. El último superado es el definitivo. Es como en el salto de altura.

---Tres puntualizaciones:

--una significación estadística sólo permite establecer una relación de causalidad si se trata de un estudio experimental

--Una diferencia estadísticamente importante no quiere decir de forma automática que lo encontrado tenga importancia práctica. Eso lo dirán las circunstancias.

--si hay significación estadística, hay que buscar siempre la posible existencia de **factores de confusión**. Así un estudio puede sugerir que los alcohólicos tienen un riesgo alto de padecer cáncer de pulmón, pero resulta que casi todos los alcohólicos eran fumadores. En estos casos hay que estratificar en subgrupos del presunto ·confundidor”

---Errores de las decisiones estadísticas

-Como un α% de los PaM caen en la zona de rechazo, aunque H_0 sea verdadera, todo rechazo de H_0 conlleva un riesgo de error, el **ERROR TIPO I**, que es el que se comete cuando se acepta H_1 siendo H_0 verdadera. Podría decirse que es un FALSO POSITIVO. Su riesgo es α. Este riesgo lo fijamos nosotros y es por tanto conocido. por consenso el máximo riesgo que se admite es de 0,05 (ó 5%). Si no se dice otra cosa se acepta ese valor de α.. El error tipo I puede ser disminuido aumentando el tamaño muestral.

-El **ERROR TIPO II** es el que se comete al no rechazar H_0 , siendo H_1 verdadera. Equivale a un falso negativo. El riesgo de cometerlo se llama β (beta) y no lo conocemos exactamente, aunque hay métodos para estimarlo, que no veremos aquí. El problema es que si queremos disminuirlo, aumentamos α , y viceversa. Las fórmulas para el tamaño muestral tienen en cuenta esta circunstancia y , asumiendo una β entre 0,05 y 0,1 .En todo caso β disminuye también aumentando el tamaño de la muestra. A $1-\beta$ se la llama potencia de una prueba estadística.

> Las decisiones estadísticas no "demuestran" nada. Sólo apoyan de una forma razonable una decisión o hecho concreto.
> Aceptar H_1 equivale a decir con un pequeño riesgo de error (α) que H_0 es falsa.
> No rechazar H_0 no quiere decir que sea verdadera, sólo que no ha podido ser rechazada (el riesgo β acecha...)

---Grado de significación

Se expresa por el mismo número que α, pero el concepto es ligeramente distinto. Es la probabilidad de que un resultado alcance un determinado valor cuando H_0 es verdadera. Cuantifica también la p de cometer un error tipo I. Su símbolo es p. Y se expresa como veíamos para α: $p < 0,05$ ó $p< 0,01$...

---Pruebas paramétricas y no paramétricas.

-**Las pruebas paramétricas** utilizan en sus cálculos parámetros, como media, varianza, frecuencia, porcentaje, etc.. Estas pruebas tienen unas condiciones de aplicación, que se especifican en cada prueba. Las mas frecuentes son: normalidad de la población origen, igualdad de varianzas, y tamaño adecuado. En la práctica, si la muestra es grande (≥ 30) cumple siempre. Por tanto es en las muestras pequeñas donde hay que comprobar las condiciones de aplicación. Si no las cumplen, no pueden utilizarse esas pruebas y hay que recurrir a las pruebas no paramétricas, que no tienen condiciones de aplicación y se pueden utilizar siempre. Algunas pruebas son muy robustas (como el ANOVA) y la no observancia de las condiciones de aplicación no altera sustancialmente la decisión estadística, por lo que casi nunca se tienen en cuenta.

-**Las pruebas no paramétricas** se basan en la comparación de los datos aislados y en su ordenación según el criterio propio de cada test.. A igualdad de tamaño muestral son menos eficientes que las prueba paramétricas, por lo que siempre que sea posible se deben usar éstas.

Recordatorio

ZONA DE ACEPTACION Y ZONA DE RECHAZO

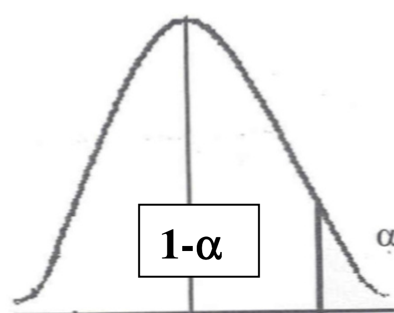
las pruebas estadísticas prueban la hipótesis nula H_0, que puede rechazarse o no rechazarse

la zona de no rechazo corresponde a **1-α**

la zona de rechazo correponde a **α** , y puede ser única, en un solo lado de la campana (pruebas unilaterales) o doble, en ambos lados de la campana (pruebas bilaterales).

Al rechazar H_0 se acepta H_1 y en las pruebas bilaterales (casi todas) hay que dar el sentido del rechazo

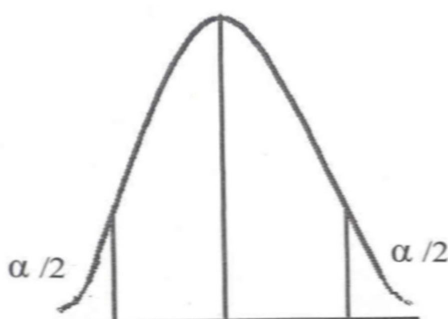
El no rechazo de H_0 no prueba que H_0 sea verdadera, sólo que no puede rechazarse (algo así como una absolución por falta de pruebas, que no afirma que el acusado sea inocente, sino que no hay pruebas para considerarlo culpable)



Prueba unilateral

H_1 es única

A > B



Prueba bilateral

H_1 es doble:
A > B
B < A

En principio las pruebas son bilaterales, y si no se dice otra cosa hay que entender que la prueba es bilateral. Interesa cualquier tipo de diferencias o relaciones.

Decisión estadística	p acertar	p no acertar	¿conocido?	< riesgo si
no rechazo H_0	1 - β = potencia	β	no	> N < β
rechazo de H_0 y aceptación H_1	1 - α = nivel de confianza	α = nivel de significación	sí 0,05 ó menos	> N < α

Tema 14 : Estimación de parámetros. Pruebas de conformidad.

Estimación de parámetros

A partir de una muestra nunca podemos saber exactamente el valor de los parámetros poblacionales, pero sí podemos estimarlos de una forma razonable con un pequeño margen de error, que podemos medir.

La mejor estimación de un parámetro de la población a partir de una muestra es 1) **el parámetro de la muestra**, sólo si la muestra es grande 2) el **intervalo de confianza** (I ó IC) del parámetro de la muestra en todos los casos (sea grande o pequeña).

Hay pues 2 tipos de estimación:

--la **estimación puntual**, que sólo es posible si la muestra es grande: PaP ≈ PaM

--la **estimación por intervalo**, que siempre es posible :

PaP ≈ PaM ± E ó PaP ≈ ∈ (PaM-E ÷ PaM+E)

siendo **E** el error muestral: **E = c*e** (muestra grande) ó **E =** $\mathbf{t_{n-1}}$***e** (muestra pequeña).

e es el **error estándar**, que como ya hemos visto es la desviación estándar de la media de los parámetros muestrales hallados en múltiples muestras obtenidas de una población (no confundir con la desviación estándar de una muestra). Es posible calcularlo ya a partir de una sola muestra.

para un porcentaje o proporción:

$$e = \sqrt{\frac{pq}{N}}$$

para una media:

$$e = \frac{s}{\sqrt{N}}$$

*** Estimación por intervalo de un porcentaje o proporción :

es el intervalo de confianza del porcentaje o proporción de la muestra

a) muestra grande

$$I_p = p \pm c\sqrt{\frac{pq}{N}}$$

b) muestra pequeña

$$I_p = p \pm t_{(n-1)}\sqrt{\frac{pq}{N}}$$

Ejemplo 1: En una muestra de 100 estudiantes de la Facultad F el 20% tienen ordenador portátil. Estimar el porcentaje de la población que tendrá ordenador portátil
-estimación puntual (es muestra grande) : 20%
-estimación por intervalo : $\mathbf{I_p = 20 \pm 1{,}96\sqrt{\frac{20*80}{100}} = 20 \pm 7{,}84 = \in (12{,}2 \div 27{,}8)\%}$
Ejemplo 2:Se hace el mismo estudio pero en una muestra de 25 alumnos y lo tienen el 20%.
-estimación puntual : no es posible, pues la muestra es pequeña
-estimación por intervalo: $\mathbf{I_p = 20 \pm 2{,}064\sqrt{\frac{20*80}{25}} = 20 \pm 16{,}51 = \in (3{,}49 \div 36{,}5)\%}$
Las fórmulas aquí expuestas son las más sencillas y suficientes para la práctica. En determinados casos puede ser necesario un cálculo más exacto (aunque sigue siendo aproximado) para el que se precisan programas estadísticos, dada su complejidad.. EPITABLE da los IC calculados por el método "cuadrático de Fleiss", el "binomial exacto" y el de la "p media (mid-p)" . En los dos ejemplos anteriores los límites son:

	cuadrático de Fleiss	binomial exacto	p media
Ejemplo 1	$\in$(12,9÷ 29,4)	$\in$(12,7 ÷ 29,4)	$\in$(12,9 ÷ 29,2)
Ejemplo 2	$\in$(7,60 ÷ 41,3)	$\in$(6,83 ÷40,7)	$\in$(7,72 ÷ 38,9)

*** Estimación por intervalo de una media:
es el intervalo de confianza de la media de la muestra

a) muestra grande $$I_{\bar{X}} = \bar{X} \pm c\frac{s}{\sqrt{N}}$$

c) **muestra pequeña** $$I_{\bar{X}} = \bar{X} \pm t_{(n-1)}\frac{s}{\sqrt{N}}$$

En estas fórmulas lo que sigue al signo ± es **E** y lo que sigue a c ó t es **e**

Ejemplo 1: En una muestra de tamaño 100 la media vale 33 y la desviación estándar 10.
-estimación puntual: 33
--estimación por intervalo: $\mathbf{I_{\bar{X}} = 33 \pm 1{,}96\frac{10}{\sqrt{100}} = 33 \pm 1{,}96 = \in (31{,}04 \div 34{,}96)}$
Ejemplo 2: Como en el ejemplo anterior, pero con una muestra de 25
-estimación puntual : no es posible, pues la muestra es pequeña
-estimación por intervalo: $\mathbf{I_{\bar{X}} = 33 \pm 2{,}064\frac{10}{\sqrt{25}} = 33 \pm 4{,}13 = \in (28{,}87 \div 37{,}13)}$

*** Estimación por intervalo de un coeficiente de correlación
Es su intervalo de confianza. Su cálculo exacto es bastante complicado.
Veremos dos métodos:
1) gráfico de David
2) método de Zr (transformación de Fisher)

---**El gráfico de David** es un método muy sencillo, que no precisa cáculos, pero su estimación es bastante burda. Se busca en la parte superior el valor de r y se une por una linea vertical imaginaria con el de la parte inferior; se marcan los puntos en que esa linea corta a las dos del tamaño

muestral; esos puntos trasladados horizontalmente a la escala lateral dan los límites del IC de r. Dada el poco detalle del gráfico hay que hacer interpolaciones. Ver el gráfico en la página 14-6 Para una r de 0,600 y un tamaño muestral de 50 el límite inferior está en 0,400 y el superior en 0,750 .**ICr ≈ ∈ (0,400 ÷ 0,750)**

---- **por Zr** : Es la transformación de Fisher, que sigue la distribución normal.

$$-- \ \mathbf{Zr} = \left(\frac{1}{2}\ln\frac{1+r}{1-r}\right) \pm \frac{c}{\sqrt{N-3}}$$

aquí se abren dos opciones:

1) utilizar la tabla de Zr para leer los límites del intervalo (ver la tabla en la página 14-7):

--el resultado se redondea a 2 decimales; si la muestra es pequeña, se toma t_{n-3} en vez de c

--se busca en la tabla a que valores de r corresponden estas dos Zr; son los límites del IC. En el ejemplo las Zr valen 0,41 y 0,98, a las que corresponden en la tabla , redondeando a 3 decimales, valores de r = 0,388 y 0,753 ; **ICr = ∈ (0,388 ÷ 0,753)**

2) utilizar una fórmula, que invierte la transformación inicial (cálculo exacto):

Para cada valor de Zr : $$\mathbf{r := \frac{e^{2Zr} - 1}{e^{2Zr} + 1}}$$

En el ejemplo se obtiene: $\mathbf{IC_r = \in (0,386 \div 0,753)}$

El gráfico de Davis nos ha dado una buena aproximación

Otro ejemplo: r = 0,400 ; N = 50 ---Davis: **ICr = ∈ (0,140 ÷ 0,600)** ---Zr . Las Zr valen 0,14 y 0,71 y por tanto **ICr = ∈ (0,139 ÷ 0,611)** ---El cálculo exacto da **ICr = ∈ (0,137÷ 0,610)**

* * * Pruebas de conformidad

Sirven para comprobar si una muestra puede proceder de una población determinada. Contestan a las preguntas: ¿puede proceder una muestra de media (o porcentaje) tal de una población de media (o porcentaje) cual? ; ¿es conforme la muestra con lo esperado para la población?...

Fundamento estadístico

Ver si el parámetro de la muestra está dentro del intervalo de probabilidad de la población.

H_0 : no hay diferencias significativas entre muestra y población; por tanto **sí** puede proceder.

H_1: hay diferencias significativas entre muestra y población; por tanto **no** puede proceder.

Si no se dice otra cosa, se toma como único nivel de significación el de 0,05.

Técnica

Es la habitual en los procesos de contraste:

---se aplica la fórmula adecuada, que depende del tipo de variable y su tamaño. Al resultado lo llamamos Z.

---comparamos Z , tomado en su valor absoluto, |Z| ,con el patrón de referencia:

-si Z es menor: no se rechaza H_0 . Se concluye que sí puede proceder, que es conforme...

-si Z es igual o mayor que el patrón de referencia: se rechaza H_0 y se acepta H_1; es decir, se concluye que no puede proceder, que no es conforme...

Fórmulas

Veremos tres: las correspondientes a la conformidad de una proporción o porcentaje, la conformidad de frecuencias y la conformidad de una media.

1) **proporción o porcentaje**

$$\mathbf{Z} = \frac{\mathbf{p_m - p_p}}{\sqrt{\frac{\mathbf{p_p * q_p}}{\mathbf{N}}}}$$

Valoración

- ♦ *muestra grande : por la DN*
- ♦ *muestra pequeña: se multiplican **p** y **q** de la población por N*
  - *--si ambos productos son ≥ 5 (ó 500, si es %) : por la DN*
  - *--si alguno de ellos es < 5 (ó 500) : por t_{n-1}*

2) **frecuencias**

usar la fórmula de contraste nº 3 (ver página 16-4)

Valoración por χ^2 con g.l. = nº de modalidades - 1

3) **media**

$$\mathbf{Z} = \frac{(\bar{\mathbf{X}}_m - \bar{\mathbf{X}}_p)\sqrt{\mathbf{N}}}{\mathbf{s}}$$

Valorar por DN, si es muestra grande; si es pequeña por t_{n-1}

En la conformidad de medias hay que tomar la s de la población, si es conocida. Si no lo es, se toma la s de la muestra, que es su mejor estimación.

Ejemplos:

1- La enfermedad A se sabe que tiene una mortalidad del 25%. Observamos una epidemia de 80 casos, de los que fallecen 24. ¿Es aún una epidemia" normal" o es más grave?

Solución:

Para aplicar la fórmula necesitamos calcular p_m y q_p , pues el resto ya lo conocemos.

p_m es el % de defunciones: 24*100/80 = 30%. q_p = 100-25 = 75%

$$\mathbf{Z} = \frac{\mathbf{30 - 25}}{\sqrt{\frac{\mathbf{25 * 75}}{\mathbf{80}}}} = \mathbf{1,03}$$

H_0: es conforme, es una epidemia "normal" al ser N>30, valoramos por $c_{0,05}$ = 1,96 . z < $c_{0,05}$; por tanto no se puede rechazar H_0 . Lo observado está dentro de lo esperado, es conforme, las diferencias observadas se explican por las variaciones del azar. Y contestando a la pregunta: No podemos rechazar la hipótesis de que se trata de una epidemia "normal".

2- Como todo porcentaje puede ser transformado en frecuencia y viceversa, este ejercicio se puede resolver contrastando las frecuencias observadas (O) y las esperadas (E), utilizando la fórmula de contraste nº 3

Fórmula nº 3

Si todas las $E \geq 5$:

$$Z = \sum \frac{(O-E)^2}{E}$$

Si alguna E<5 pero ≥ 3:

$$Z = \sum \frac{(|O-E| - 0,5)^2}{E}$$

Si alguna E<3: no aplicable

Valoración: por $\chi^2_{(f-1)(k-1)}$

si no es aplicable por ser E<3, hay que utilizar la p exacta de Fisher

	O	E
Muertos	24	20
Vivos	56	60
TOTAL	80	80

Z = $(24-20)^2/20 + (56-60)^2/60$ =1,07, que es menor que $\chi^2(1 , 0'05)$ = 3,84. Por tanto no se puede rechazar H_0 y se llega a la misma conclusión

3-: ¿Puede proceder una muestra de 20 personas con un número de fumadores de 10 de una población de fumadores del 45%?
--- Problema de conformidad entre la proporción o porcentaje observado en una muestra y lo esperado en una población. H_0 : no hay diferencias significativas entre muestra y población, sí puede proceder la muestra de esa población, hay conformidad. Pm=(10/20)*100=50% Pp= 45% Qp=55% N=20

Z = (50-45)/√(45*55)/20 = 0'45 .

Es muestra pequeña : como N*Pp y N*Qp >500 , se valora por c de la DN : Z < c0'05 (=1,96) y por tanto no se puede rechazar H_0 : **Sí puede proceder**

---- También se puede resolver contrastando frecuencias, las observadas en una muestra y las teóricas correspondientes a una población. H_0 : no hay diferencias significativas entre muestra y población, sí puede proceder la muestra de esa población, hay conformidad.

	O	E
Fumad.	10	9
No Fum.	10	11
TOTAL	20	20

Z = $(10-9)^2$ /9 + $(10-11)^2$ /11 = 0'20 , que es menor que X^2 (1, 0'05) = 3.84 y por tanto no se puede rechazar H_0; **SÍ** puede proceder la muestra de esa población.

4-: Un Laboratorio Farmacéutico afirma que las tabletas XYZ calman el dolor de estómago durante por lo menos 4 horas en una proporción de 0'85. Para comprobarlo se hace una experiencia con 20 personas enfermas, elegidas al azar. El resultado es positivo en 12 pacientes. ¿Está este resultado de acuerdo con lo afirmado por el Laboratorio?
---- Problema de conformidad entre la proporción o porcentaje observado en una muestra pequeña y lo esperado en una población. H_0 : no hay diferencias significativas entre muestra y población, sí puede proceder la muestra de esa población, hay conformidad, el resultado está de acuerdo con lo afirmado por el Laboratorio.

Pm=(12/20)*100=60% Pp=85% Qp=15% N=20

Z = (60-85)/√(85*15)/20 = -3'13

Como es muestra pequeña y N*Qp <500 se valora por t(19,0'05)=2'093
|Z| > t y por tanto se rechaza H0 a ese nivel de significación. La muestra no es conforme con la población: el resultado obtenido no está de acuerdo con lo afirmado. Sentido: Hemos obtenido un resultado peor.

---- También se puede resolver contrastando frecuencias, las observadas en una muestra y las teóricas correspondientes a una población. H_0 : no hay diferencias significativas entre muestra y población, sí puede proceder la muestra de esa población, hay conformidad.

	O	E
Calma	12	17
No calma.	8	3
TOTAL	20	20

Como un valor E es <5, pero mayor de 3 : **Z = $(|12-17|-0'5)^2$ /17 + $(|8-3|-0'5)^2$ /3 = 7'94**, que es mayor que X^2 (1, 0'05) = 3.84 y por tanto, igual que antes, se rechaza H_0 .

5-: Un Laboratorio farmacéutico declara que sus tabletas ABC contienen 100 mg de producto activo, con una varianza de 100. Hacemos una prueba con 36 tabletas tomadas al azar y encontramos una media de 95 mg con s = 12. ¿Contradice este resultado lo afirmado por el Laboratorio?
--- Problema de conformidad entre la media aritmética de una muestra y la de la población. H_0: esa muestra puede proceder de la población, es conforme con ella, no hay diferencias significativas entre ambas...
Valoración por la c correspondiente a 0'05, que vale 1'96 (es muestra grande)

|Z| = ((95-100)*√36) /10 = -3

|-3 | = 3 > 1'96 , luego se rechaza H_0: : no hay conformidad, la muestra no puede proceder de esa población, los resultados contradicen lo afirmado por el fabricante. Sentido: hay menos principio activo
Ya que se conoce, se toma la s de la población (s=√100 = 10) y no la de la muestra.

Anexos:

Gráfico de David

Intervalos del 95 por 100 para el coeficiente de correlación *

(De F. N. DAVID, *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples,* The Biometrika Office. Londres, 1938.)

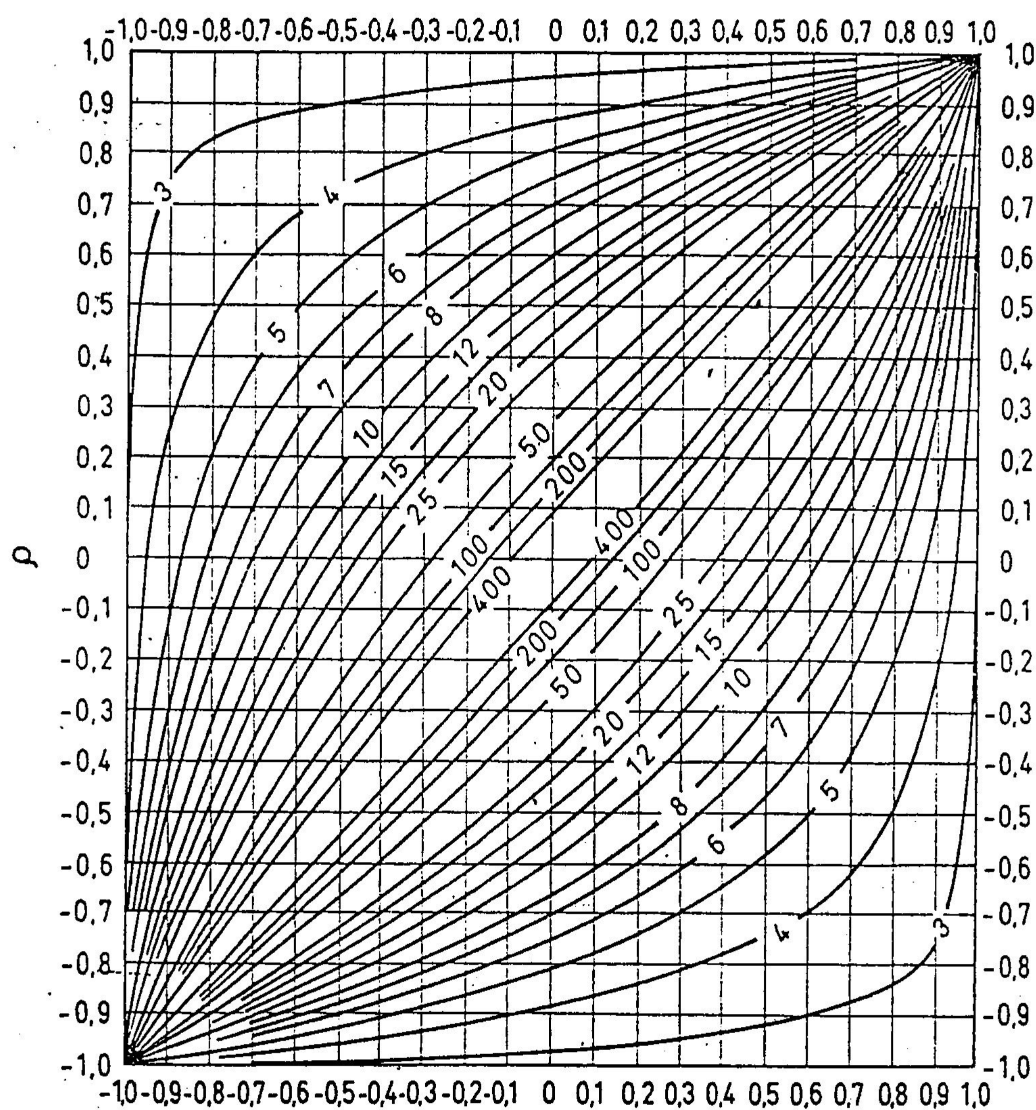


Tabla de Zr

Valores de r para distintos valores de z_r

(De *Statistical Methods for Research Workers*, por R. A. FISCHER, Oliver and Boyd. Edimburgo.)

z_r	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0699	0,0798	0,0898
0,1	0,0997	0,1096	0,1194	0,1293	0,1391	0,1489	0,1587	0,1684	0,1781	0,1878
0,2	0,1974	0,2070	0,2165	0,2260	0,2355	0,2449	0,2543	0,2636	0,2729	0,2821
0,3	0,2913	0,3004	0,3095	0,3185	0,3275	0,3364	0,3452	0,3540	0,3627	0,3714
0,4	0,3800	0,3885	0,3969	0,4053	0,4136	0,4219	0,4301	0,4382	0,4462	0,4542
0,5	0,4621	0,4700	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299
0,6	0,5370	0,5441	0,5511	0,5581	0,5649	0,5717	0,5784	0,5850	0,5915	0,5980
0,7	0,6044	0,6107	0,6169	0,6231	0,6291	0,6352	0,6411	0,6469	0,6527	0,6584
0,8	0,6640	0,6696	0,6751	0,6805	0,6858	0,6911	0,6963	0,7014	0,7064	0,7114
0,9	0,7163	0,7211	0,7259	0,7306	0,7352	0,7398	0,7443	0,7487	0,7531	0,7574
1,0	0,7616	0,7658	0,7699	0,7739	0,7779	0,7818	0,7857	0,7895	0,7932	0,7969
1,1	0,8005	0,8041	0,8076	0,8110	0,8144	0,8178	0,8210	0,8243	0,8375	0,8306
1,2	0,8337	0,8367	0,8397	0,8426	0,8455	0,8483	0,8511	0,8538	0,8565	0,8591
1,3	0,8617	0,8643	0,8668	0,8693	0,8717	0,8741	0,8764	0,8787	0,8810	0,8832
1,4	0,8854	0,8875	0,8896	0,8917	0,8937	0,8957	0,8977	0,8996	0,9015	0,9033
1,5	0,9052	0,9069	0,9087	0,9104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9202
1,6	0,9217	0,9232	0,9246	0,9261	0,9275	0,9289	0,9302	0,9316	0,9329	0,9342
1,8	0,9354	0,9367	0,9379	0,9391	0,9402	0,9414	0,9425	0,9436	0,9447	0,9458
1,7	0,9468	0,9478	0,9498	0,9488	0,9508	0,9518	0,9527	0,9536	0,9545	0,9554
1,9	0,9562	0,9571	0,9579	0,9587	0,9595	0,9603	0,9611	0,9619	0,9626	0,9633
2,0	0,9640	0,9647	0,9654	0,9661	0,9668	0,9674	0,9680	0,9687	0,9693	0,9699
2,1	0,9705	0,9710	0,9716	0,9722	0,9727	0,9732	0,9738	0,9743	0,9748	0,9753
2,2	0,9757	0,9762	0,9767	0,9771	0,9776	0,9780	0,9785	0,9789	0,9793	0,9797
2,3	0,9801	0,9805	0,9809	0,9812	0,9816	0,9820	0,9823	0,9827	0,9830	0,9834
2,4	0,9837	0,9840	0,9843	0,9846	0,9849	0,9852	0,9855	0,9858	0,9861	0,9863
2,5	0,9866	0,9869	0,9871	0,9874	0,9876	0,9879	0,9881	0,9884	0,9886	0,9888
2,6	0,9890	0,9892	0,9895	0,9897	0,9899	0,9901	0,9903	0,9905	0,9906	0,9908
2,7	0,9910	0,9912	0,9914	0,9915	0,9917	0,9919	0,9920	0,9922	0,9923	0,9925
2,8	0,9926	0,9928	0,9929	0,9931	0,9932	0,9933	0,9935	0,9936	0,9937	0,9938
2,9	0.9940	0,9941	0,9942	0,9943	0,9944	0,9945	0,9946	0,9947	0,9949	0,9950
3,0	0,9951									
4,0	0.9993									
5,0	0,9999									

Tema 15 : PRUEBAS DE CONTRASTE DE VARIABLES

Veremos únicamente el contraste (comparación o relación) de **dos** variables. Para ello se dispone de 15 pruebas o tests estadísticos, que se eligen en función de la naturaleza de las variables, del nº de modalidades de las variables cualitativas (CL) y del tipo de datos (independientes o apareados). Cuando los datos son independientes, las fórmulas para problemas de comparación y relación son las mismas; si los datos son apareados, son distintas. Para la mayoría de las situaciones se dispone además de la prueba paramétrica, que es la de elección, de otra no paramétrica.
La siguiente tabla sirve de guía para elegir la prueba adecuada. Se puede entrar en ella por dos sitios: la primera columna (variables) y la cuarta columna (contraste de...)

PRUEBAS DE CONTRASTE DE VARIABLES

Variables	Datos	Modalidades	Contraste de...	Fórmula nº
CL y CL	Independ.	2	**Proporciones o porcentajes**	1
		2	**Frecuencias**	2
		3 ó más	**Frecuencias**	3
	Aparead.	2	<u>Prueba de comparación</u> ■ **proporciones ó %** ■ **frecuencias** ------------------------- <u>prueba de relación</u> ■ **proporciones ó %** ■ **frecuencias**	 4 5 ---------------------- 1 2
CL y CT	Independ.	2	**<u>dos medias</u>** ■ paramétrico ■ no paramétrico	 6 7 Mann-Whitney
		3 ó más	**<u>k medias</u>** ■ paramétrico ■ no paramétrico	 8 ANOVA-1 9 Kruskal-Wallis
	Aparead.	2	**<u>2 medias</u>** <u>- prueba de comparación</u> ■ paramétrico ■ no paramétrico ------------------------- <u>- prueba de relación</u> como si fuera CT y CT	 10 11 Pª de los signos ---------------------- 14 ó 15
		3 ó más	**<u>k medias</u>** (prueba de comparación) ■ paramétrico ■ no paramétrico	 12 ANOVA-2 13 Test de Friedman
CT y CT	Todos		**<u>Coeficiente de correlación</u>** ■ paramétrico ■ no paramétrico	 14 (r de Pearson) 15 (r de Spearman)

PASOS EN EL CONTRASTE DE VARIABLES

1) encontrar la fórmula adecuada

Hay dos caminos:

***empezar por la primera columna:
---reconocer las variables (y las modalidades en las CL)
---¿datos independientes o apareados?
→ pasar por el tipo de contraste (de p ó % , de f , de medias...) al nº de fórmula

***empezar por la columna central del tipo de contraste:
---¿qué me piden que contraste,
p, %, medias....?
una vez identificado:
→ la 1ª columna y seguir como arriba

2) definir H_0 : no hay diferencias o relación entre las variables contrastadas

3) ¿hay condición de aplicación?

si la hay, ¿se cumple?

4) aplicar la fórmula : obtenemos un resultado al que genéricamente llamamos Z

5) comparar Z y el valor de referencia que corresponda

6) tomar la decisión estadística

---no rechazo de H_0 : Z < valor de referencia
---rechazo de H_0 y aceptación de H_1: Z ≥ valor de referencia
en este caso: --a qué nivel de significación
--sentido del rechazo

Tema 16: Contraste de dos variables cualitativas. Odds ratios.

En el contraste de dos variables cualitativas hay que ver 1) si se trata de datos independientes o apareados 2) el número de modalidades de las variables (dos o más de dos). Ya que se utilizan fórmulas distintas

A) Contraste de 2 variables cualitativas con datos independientes

Como en toda prueba con datos independientes los problemas de comparación y de relación se resuelven por las mismas fórmulas, ya que son dos formas distintas de enfocar el mismo problema..

Responden a las preguntas:

----la frecuencia (absoluta, relativa o porcentaje) de una característica ¿es similar en los grupos o muestras contrastados?.

En caso afirmativo se trata de una **prueba de comparación**. H_0 : no hay diferencias significativas entre las frecuencias contrastadas, las diferencias observadas se deben a las variaciones normales por el azar.

----¿hay relación o dependencia entre las muestras contrastadas?

En caso afirmativo es una **prueba de relación**. H_0: NO hay relación o dependencia.

Fórmulas

En función del nº de modalidades y de los datos aplicaremos una de las fórmulas siguientes:

- Cuando ambas variables tienen dos modalidades:
  *** **Fórmula nº 1** : para contraste de proporciones o porcentajes
  *** **Fórmula nº 2** : para el contraste de frecuencias absolutas
- Si una o ambas variables tienen más de dos modalidades:
  *** **Fórmula nº 3** : en la práctica sólo se utilizan frecuencias absolutas

(es más fácil utilizar porcentajes que proporciones)

Fórmula nº 1

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_0 q_0}{N_1} + \frac{p_0 q_0}{N_2}}} \qquad \text{siendo } p_0 = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$

Valoración: si N1 y N2 ≥30, por la DN

si N1 ó N2 <30

a) si p_0N_1 , q_0N_1, p_0N_2 y q_0N_2 ≥5 (ó 500 si es %) por DN

b) ***si algún producto <5 pero > 3 : por t (N_1+N_2 –2)***

si algún producto < 3 : por p exacta de Fisher

***Ejercicio 1.1

En una muestra de 100 varones encontramos un 70% de fumadores. En una muestra de 200 mujeres hay 80 fumadoras. ¿Hay diferencias importantes en el hábito de fumar entre ambos sexos?

---- Se trata de un problema de contraste entre dos variables CL (Sexo , Hábito de fumar) con dos modalidades cada una (Hombre, Mujer y Sí, No) con datos independientes. H_0 : no hay diferencias significativas entre los variables contrastadas.

Se puede resolver por la fórmula nº 1 (contraste de dos porcentajes) o por la fórmula nº 2 (contraste de 2 frecuencias). Lo haremos por ambas, pero si podemos elegir, es preferible la nº 2. En la nº 1 es mejor utilizar % que proporciones.
Empezamos por la nº 1 : Hemos medido el hábito de fumar en hombres y en mujeres.
Por el enunciado o mediante un pequeño cálculo sabemos que
$p_1=70$; $N_1 = 100$; $p_2 = 40$; $N_2 = 200$; $p_0 = 50$; $q_0 = 50$

$$Z=\frac{70-40}{\sqrt{\frac{50*50}{100}+\frac{50*50}{200}}}=4{,}90$$

Como N_1 y N_2 son > 30, se valora por c de la DN
$Z > c_{0,001} = 3{,}30 \rightarrow$ rechazo de H_0 : y aceptación de H_1 al nivel de significación de 0,001. ; **p < 0,001**
Sentido: el % de hombres fumadores es más alto..
Y contestando a la pregunta: **Sí** hay diferencias importantes, los hombres fuman más

Fórmula nº 2

$$Z=\frac{N\left(a_1b_2-a_2b_1\right)^2}{N_aN_bN_1N_2}$$

tabla:

a_1	a_2	N_a
b_1	b_2	N_b
N_1	N_2	N

Condición de aplicación: todas las E ≥ 5
Valoración: por χ^2_1
Si **alguna E<5 , pero ≥ 3: usar fórmula de Yates**
Si alguna E<3: calcular p exacta de Fisher
Fórmula de Yates:

$$Z=\frac{N\left(\left|a_1b_2-a_2b_1\right|-\frac{N}{2}\right)^2}{N_aN_bN_1N_2}$$

***Ejercicio 1.2 : vamos a resolver el ejercicio anterior por la fórmula nº 2

1---Se construye la tabla de 2x2:

	Fuma	No fuma	
Hombre	70	30	*100*
Mujer	80	120	*200*
Total	*150*	*150*	300

2--- se comprueba condición de aplicación:
cumple, pues la E más baja (en a_1 y a_2) vale 50 y es > 5
3---se calcula Z

$$Z=\frac{300*(70*120-80*30)}{100*200*150*150}=24$$

4---Se valora por χ^2 con gl = 1 $Z > \chi^2\ (1,\ 0{,}001) = 10{,}83$; **p < 0,001**
Por tanto se rechaza H_0 y se acepta H_1 : hay diferencias significativas a nivel de 0,001 entre las frecuencias de fumadores en hombres y mujeres. Sentido: los hombres fuman más. **Sí** hay diferencias importantes.

***Ejercicio 1.3 En una muestra de 20 personas de la tercera edad de la ciudad A el 30% tiene un colesterol alto. En la ciudad B lo tienen el 50% de una muestra de 30. ¿Es importante esa diferencia?

----Es un problema de comparación entre dos variables CL con 2 modalidades cada una y datos independientes: COLESTEROL (alto, no alto) y CIUDAD (A , B)

H_0 : no hay diferencias significativas entre las variables contrastadas, las diferencias observadas se explican por las variaciones normales del azar.

La tabla guía nos indica que lo podemos resolver por la fórmula 1 ó la fórmula 2

Vamos a hacerlo a efectos didácticos, por ambas. Es más fácil, utilizar la nº 2.

1.3.1 Resolución por la fórmula nº 1

Por el enunciado o haciendo un pequeño cálculo se sabe que

p_1=30 ; N_1= 20 ; p_2=50 ; N_2 = 30 ; p_0 = 42 ; q_0 = 58

$$Z = \frac{30-50}{\sqrt{\frac{42*58}{20}+\frac{42*58}{30}}} = -1,4$$

Como una muestra es pequeña, hay que ver lo que valen los productos de ambas N por p_0 y q_0. Todos son > 500 (el menor: 20*42=840), por lo que la Z se valora por la DN.
|Z| < 1,96 → No puede rechazarse H_0 . p > 0,05 n.s.

Y contestando a la pregunta: la diferencia no es importante.

1.3.2 Resolución por la fórmula nº 2

1---se construye la tabla de 2x2:

		Ciudad		
		A	B	
Col	alto	6	15	*21*
	no alto	14	15	*29*
		20	*30*	50

2---cumple la condición de aplicación: la E más baja (a_1) vale 21*20/50 = 8,4 que es > 5

3---se calcula Z :

$$Z = \frac{50*(6*15-15*14)}{21*29*20*30} = 1,97$$

4---se valora por χ^2 con g l = 1 ; Z < χ^2 (1, 0,05) = 3,84 → No puede rechazarse H_0 . p > 0,05 n.s. Contestando a la pregunta : la diferencia no es importante.

***Ejercicio 1.4 En un colegio se hace una encuesta en busca de miopes. Hay 4 entre 20 chicos y 7 entre 28 chicas. Valore la afirmación: la miopía es más frecuente entre las chicas.

--- Es un problema de contraste entre dos variables CL con 2 modalidades cada una : MIOPIA (sí, no) y SEXO (chico , chica). Datos independientes. A resolver por la fórmula nº 1 ó la nº 2.

H_0 : no hay diferencias significativas entre los variables contrastadas.

---no vemos en detalle la resolución por la fórmula nº 1. p0 vale 22,9% y qo 77,1%. Se obtiene una Z = -0,406, que hay que valorar por t(46 , 0'05) = 2,014 . |Z| < t →No puede rechazarse H_0 . La afirmación no está justificada estadísticamente.

---resolución por la fórmula nº 2 :

1---construir la tabla:

		Miopía		
		Sí	No	
Sexo	Chico	4	16	*20*
	Chica	7	21	*28*
		11	*37*	48

2---Hay una E < 5 (la a1, que vale 4,6) → fórmula de Yates

3-- $$Z = \frac{48*\left(|(4*21)-(7*16)|-\frac{48}{2}\right)^2}{20*28*37*11} = 0,003$$

4--- Z<χ^2 (1, 0'05) = 3,84 →No puede rechazarse H_0. p>0,05 n.s. La afirmación no está justificada

***Ejercicio 1.5   Se estudia el efecto de la vacuna BCG en la prevención de la TBC (tuberculosis) en el pueblo X de un país en vías de desarrollo. Hay 10 enfermos entre 70 vacunados y 80 enfermos entre 120 no vacunados. ¿Tiene la vacuna efecto preventivo?

--Es un problema de contraste de 2 variables CL con dos modalidades cada una y datos independientes: BCG (sí , no) y TBC (sí , no). A resolver por la fórmula nº 1 o la nº 2. Lo haremos por la nº 2, pues es más fácil y por tanto preferible.

		BCG SI	BCG NO	
TBC	SI	10	80	90
	NO	60	40	100
		70	120	190

Cumple condición de aplicación: todas las E ≥5
Z= 48'66 >χ 2 (1 , 0'001)=10'83 y por tanto se rechaza H_0 al nivel de significación de 0'001: Sí hay diferencias. **p < 0,001** .
Sentido: los vacunados enfermas menos. "La vacuna tiene efecto preventivo".

B) Contraste de 2 variables CL con datos independientes y 3 ó + modalidades

Fórmula nº 3

Si todas las E ≥ 5 :

$$Z = \sum \frac{(O - E)^2}{E}$$

Si alguna E < 5 pero ≥ 3 :

$$Z = \sum \frac{(|O - E| - 0,5)^2}{E}$$

Si alguna E < 3 : no aplicable

Valoración : por $\chi^2_{(f-1)(k-1)}$

***Ejercicio 1.6 Se realiza un experimento de germinación con 3 tipos de semillas en un terreno abonado con la sal S al 5%. De 25 semillas de la especie A germinan 15, de 30 de la B germinan 25 y lo hacen 19 de las 25 de la especie C. ¿Se comportan las especies de forma distinta?

------Problema de comparación de dos Vbles. CL : ESPECIE, con 3 modalidades - A, B y C- y GERMINACION, con 2 modalidades –sí , no. Datos independientes.
A resolver por la fórmula nº 3.
H_0 : no hay diferencias significativas ; germinan de forma similar

		Germinación SI	Germinación NO	
Esp.	A	15	10	25
	B	25	5	30
	C	19	6	25
		59	21	80

Se calculan las E y se añaden a la tabla . Cumple la condición de aplicación: todas las E ≥5

Germinación

Esp.		SI	NO
	A	**15** ; *18'43*	**10** ; *6'56*
	B	**25** ; *22'12*	**5** ; *7'87*
	C	**19** ; *18'43*	**6** ; *6'56*

Se aplica la fórmula nº 3 : Z=3'93 < χ^2 (2 ; 0'05)= 5,99
No se puede rechazar H_0. , p > 0,05
"No , el comportamiento es similar"

***Ejercicio 1.7 En 250 personas, elegidas al azar, encontramos las siguientes combinaciones de color de ojos y de pelo : (A=azul, G=gris, N=negro, R=rubio, C=castaño). En 65 A+R, en 20 A+C, en 8 A+N, en 32 G+R, en 40 G+C, en 30 G+N, en 5 N+R, en 10 N+C y en 40 N+N
¿Hay relación entre el color del pelo y el de los ojos?

Es un problema de contraste entre dos variables CL:
- COLOR OJOS con 3 modalidades (A, G y N)
- COLOR PELO con 3 modalidades (R, C y N)

y datos independientes, que se resuelve por la fórmula nº 3

H_0: no hay relación entre el color de los ojos y el color del pelo.

1) construir una tabla de 3x3:

PELO

OJOS		R	C	N	
	A	65	20	8	93
	G	32	40	30	102
	N	5	10	40	55
		102	70	78	250

2) calcular los E de cada casilla. (= total de su fila * total de su columna / total general). Vemos que todos son ≥5 y por tanto se cumple la condición de aplicación. Completamos la parte de la tabla que nos interesa, añadiendo al lado de los valores observados, los esperados (E). Los valores esperados son los que se deberían encontrar si no hubiera relación entre las variables, es decir, si H_0 fuera verdadera.

PELO

OJOS		R	C	N
	A	**65** ; *37'94*	**20** ; *26'04*	**8** ; *29'02*
	G	**32** ; *41'62*	**40** ; *28'56*	**30** ; *31'82*
	N	**5** ; *22'44*	**10** ; *15'4*	**40** ; *17'16*

3) aplicar la fórmula nº 3 : $\mathbf{Z = \sum \frac{(O - E)^2}{E}}$

Z = 19'30 + 1'40 + 15'23 + 2'22 + 4'58 + 0'10 + 13'55 + 1'89 + 30'40 = **88'67**

4) Z > $\chi 2$ (4 ; 0'001)=18'47 y por tanto se rechaza H_0 y se acepta H_1: hay relación entre el color de ojos y pelo al nivel de significación < 0'001. **p < 0,001** . Sentido: (lo vemos comparando las O y las E, nos lo dan los sumandos de Z) : los ojos negros se asocian con el pelo negro y, en menor medida, los ojos azules con el pelo rubio.

C) Contraste de 2 variables cualitativas con datos apareados

Veremos únicamente el caso de que cada variable tenga dos modalidades. Cada individuo proporciona dos datos, forma parte de ambos grupos.
Al igual que en el caso de datos independientes se plantean dos tipos de problemas:
----de comparación: ¿las frecuencias o porcentajes observados son similares en ambas muestras?

H_0: son similares, no hay diferencias significativas, las observadas se deben al azar
----de relación: ¿las variables están relacionadas entre sí?. ¿Hay dependencia entre ellas?
H_0: no hay relación o dependencia

Al ser los datos apareados, comparación y relación son dos cosas distintas, que deben ser resueltas de forma distinta, con fórmulas distintas. Para los problemas de comparación veremos dos fórmulas nuevas: la nº 4 y la nº 5. Para los problemas de relación se usan las ya vistas: nº 1 y nº 2

Pruebas de comparación

Se construye siempre una tabla de 2x2 , de forma un poco distinta a lo visto anteriormente ("se entrelazan" las variables ; los ejemplos mostrarán cómo). Sólo se tienen en cuenta los datos discordantes, aquellos en que no coinciden las variables: a uno se le llama N_1 y al otro N_2 , a la suma de ambos N

fórmula nº 4 : contraste de proporciones (si se utilizan % hay que dividir por 100)

$$Z = (p_1 - 0,5)\sqrt{4N}$$

siendo **$N = N_1 + N_2$; $N_1 = n^o$ de A+ B- ; $N_2 = n^o$ de A- B+ ; $p_1 = \frac{N_1}{N}$**

¡esta N no es la N de la tabla!

***Valoración*: si $N \geq 10$ por DN ; si <10 pero ≥ 5 por t_{N-1}; si <5: p Fisher**

fórmula nº 5 : contraste de frecuencias (más sencilla que la anterior)

los símbolos son los mismos de la fórmula nº 4

Si $N \geq 10$: $$Z = \frac{(N_1 - N_2)^2}{N}$$

Si $N<10$ y ≥ 5 : $$Z = \frac{(|N_1 - N_2| - 1)^2}{N}$$

***si* N<5 : p exacta de Fisher**

***Valoración* : por χ_1^2**

Ejercicio 2.1 En el diagnóstico de la enfermedad F se utilizan los análisis A y B. Aplicamos ambos análisis a 100 enfermos. Hay un 30% de resultados positivos con A y un 20% con B. Una cuarta parte de los positivos a B fueron negativos a A. En un 65% ambas pruebas fueron negativas. ¿Cual de los dos análisis es mejor?

---Es un problema de comparación entre 2 Vbles. CL con dos modalidades cada una y datos apareados: ANALISIS (A , B) y RESULTADO (+ , -)

Si no se ve claro que es un problema de comparación, hay que preguntarse: ¿que me piden? ¿que averigüe si los análisis diagnostican igual o uno es mejor que otro (comparación) o si los resultados de uno están relacionados con los del otro (relación)?

H_0 : no hay diferencias significativas entre los variables contrastadas. Diagnostican igual

1---construimos la tabla. Nos dan los datos de una forma un tanto enrevesada, pero con un poco de reflexión es fácil hacerlo:

		A		
		+	-	
B	+	15	5	*20*
	-	15	65	*80*
		30	*70*	100

Los datos discordantes son 15 y 5. Por tanto N_1 =15 y N_2 = 5

---2.1.1 resolución por fórmula nº 4

N_1=15 , N_2=5 , N=20 , p_1=15/20 = 0,75 $\mathbf{Z = (0,75 - 0,5) * \sqrt{4 * 20} = 2,24}$

$Z > c_{0,05}$=1,96 , por lo que se rechaza H_0 y se acepta H_1 al nivel de significación de 0,05.
p<0,05 Sentido: el análisis A es positivo con más frecuencia que B.
Contestando a la pregunta: sí, A es mejor.

---2.1.2 resolución por la fórmula nº 5

$\mathbf{Z = \frac{(15-5)^2}{20} = 5}$ $> \chi^2$ (1 , 0'05) = 3,84 → rechazo de H0 y aceptación de H1 a ese nivel de significación. **p < 0,05** . La misma conclusión que antes.

Prueba de relación

Como ya hemos visto en la página 16-5 , estos problemas se resuelven como en el caso de datos independientes por las fórmulas 1 ó 2. Y por tanto se tienen en cuenta todos los valores de la tabla.

Ejercicio 2.1.3 ¿Están relacionados los análisis del ejercicio anterior?
Está claro por la pregunta que se trata de un problema de relación. Entre dos variables CL con dos modalidades cada una y datos apareados.
Veamos la resolución por la fórmula nº 2 :
H_0 :no hay relación significativa ; no hay dependencia
Cumple la condición de aplicación: todas las E ≥ 5

$$\mathbf{Z = \frac{100*(15*65 - 15*5)^2}{20*80*70*30} = 24,11}$$

$Z > \chi^2$ (1, 0'001)=10,83 → rechazo de H_0 a ese nivel de significación y aceptación de H_1:hay una relación significativa. $p < 0,001$
Sentido: la relación es positiva

Si se aplica la fórmula nº 1, se obtiene una Z = 4,91 , que es mayor que la $c_{0,001}$ = 3,30, lo que lleva a las mismas conclusiones.

Ejercicio 3 Se prueban dos avisadores de radar, X e Y, colocados ambos en 33 vehículos, que pasan ante un radar. El X avisó en 23 casos, el Y en 25 y en 5 ocasiones no avisó ninguno. ¿Es el Y de más confianza? ¿Hay dependencia entre ellos?

Nos plantean un problema de comparación y otro de relación.

Problema de comparación (resuelto por fórmula nº 5):
Es un problema de comparación entre 2 Vbles. CL con 2 modalidades cada una y datos apareados: AVISADOR (X - Y) y AVISO (sí – no).
H_0: no hay diferencias significativas entre las frecuencias o porcentajes de las variables contrastadas, ambos aparatos avisan igual, son de igual confianza

		X		
		SI	NO	
Y	SI	20	5	25
	NO	3	5	8
		23	10	33

Sólo interesan los datos discordantes : 5 y 3 : N1=5 , N2=3 , N=8
Como N está entre 5 y 10 se aplica la fórmula nº 5 corregida:
$Z = (|5-3|-1)^2 / 8 = 0'125$, a valorar por $\chi 2$ (1, 0'05) : $Z < \chi 2$
y por tanto no se puede rechazar la hipótesis nula.
Conclusión: avisan igual .... El Y no es de más confianza

Problema de relación (a resolver por la fórmula 1 ó 2)
En ambos casos se comprueba que no cumplen la condición de aplicación.
Si elegimos la fórmula nº 1 : N_1=23 , N_2=10 , p_1=20/23=0,8696 , p_2=5/10=0'5 , p_0=0'758 , q_0=0'242. al ser muestras pequeñas hay que comprobar la condición de aplicación: N2*q0=2'42 que es <3. Hay que calcular la p exacta de Fisher (pF).
Si elegimos la fórmula nº 2 : Hay una E (la que corresponde a la casilla b2) que vale 10*8/33=2,42 y también obliga a calcular la p exacta de Fisher

p exacta de Fisher (pF)

$$p_F = \sum_{a_1=a_1}^{a_1=0} \frac{N_1!N_2!N_a!N_b!}{a_1!b_1!a_2!b_2!N!}$$

nos da la p directamente; no hay que consultar tablas. Para que sea significativa debe ser < 0,05
Esta p es para prueba unilateral, que es la que se utiliza en la práctica. Para prueba bilateral, multiplicar por 2

Los programas estadísticos la calculan fácilmente y de un tirón.
Manualmente, con la ayuda de una calculadora científica se hace siguiendo estos pasos:
1) remodelar la tabla de tal forma que en a_1 quede el valor más bajo.

		X		
		SI	NO	
Y	NO	3	5	8
	SI	20	5	25
		23	10	33

2) quedando fijos Na, Nb, N1, N2 y N , se disminuye a_1 en 1 unidad y se cambian los otros valores del interior de la tabla para que las sumas marginales fijas sean correctas. Se sigue haciendo lo mismo hasta que a_1 sea 0 Así:

2	6	1	7	0	8
21	4	22	3	23	2

3) se aplica la fórmula de la pF para cada una de las tablas y al final se suman todos los resultados parciales obtenidos.
Nota: Como Na, Nb, N1, N2 y N no cambian , recomiendo calcular y dejar en la memoria Na!Nb!N1!N2!/N! .En cada tabla dividiremos este valor almacenado entre el producto $a_1!b_1!a_2!b_2!$ y así obtendremos las p parciales, que sumadas nos dan la pF

En este problema : Na!Nb!N1!N2!/N! = $6'75675^{21}$

p parciales : para a_1 =3 0'032143978
para a_1 =2 0'00382666
para a_1 =1 0'00019879
para a_1 =0 3'2411 $^{-06}$
----
pF = 0'03617267 **p<0,05**

que al ser < 0'05 se rechaza H0 y se acepta H1 : hay relación entre los avisadores, no son independientes. Sentido: bastante coincidencia en el aviso, cuando avisa uno lo suele hacer el otro.

Odds ratio (OR)

Otros nombres: razón de probabilidades, razón de desigualdades
Es el parámetro típico de los estudios caso-control (pero la OR vale para todo tipo de estudios, que queden reflejados en una tabla de 2x2)). Se comparan dos variables CL. Un grupo de individuos que presentan una características determinada (generalmente una enfermedad : casos o "afectados") se compara con otro grupo de individuos que no la presentan (controles o "no afectados") para investigar el nivel de exposición a determinados factores que podrían ser causales. A cada caso le corresponden uno o más controles, que deben ser lo más parecidos posible a los casos, excepto en la característica en cuestión.
Se parte de la hipótesis nula: la presencia de la característica no está relacionada con la exposición. El investigador determina el tamaño muestral de los dos grupos, casos y controles, pero ignora como se reparte la exposición entre ellos. La asociación entre exposición y resultado se estima por la razón de probabilidades, más conocida con el nombre de Odds Ratio (OR), que se obtiene dividiendo las probabilidades de casos y controles. Valores posibles: $0 \leq OR \leq \infty$
En vez de la exposición se pueden estudiar los resultados de un análisis en casos y controles para ver su eficacia en el diagnóstico de la enfermedad. O se puede vigilar la aparición de una enfermedad después de haber introducido una vacuna contra la misma, etc.
:

		Enfermedad		
		+	-	
Exposición o resultado	+	$\mathbf{a_1}$	$\mathbf{a_2}$	Na
	-	$\mathbf{b_1}$	$\mathbf{b_2}$	Nb
		N_1	N_2	N

Fórmulas:

a) datos independientes :
(lo más frecuente)

$$\mathbf{OR = \frac{a_1}{a_2} : \frac{b_1}{b_2} = \frac{a_1 b_2}{a_2 b_1}}$$

b) datos apareados: $\mathbf{OR = \frac{a_2}{b_1}}$ (son los datos discordantes)

Si alguna casilla vale 0 , la OR y su intervalo de confianza pueden ser incalculables. Solución : sumar 0,5 al valor de cada casilla

Si la OR es >1 , la asociación es positiva, tanto más intensa, cuanto más alta es. La exposición favorece la aparición de la enfermedad. No hay límite superior para el valor que puede alcanzar la OR. El valor de la casilla a_1 es mayor de lo esperado.
Si la OR es < 1 , la asociación es negativa, tanto más cuanto más baja sea (aunque el número siempre es positivo, ya que por la estructura de la fórmula no puede ser < 0). La exposición dificulta la aparición de la enfermedad, protege contra la misma (p. e. una vacuna eficaz). El valor de la casilla a1 es menor de los esperado.
Si la OR es = 1 , no hay asociación; la exposición no influye nada en la aparición de la enfermedad. Es la que corresponde a H_0 .

Para <u>interpretar una OR</u> se toma como referencia el significado de la casilla a_1, que generalmente es la conjunción de enfermedad + y exposición +. Si se cambia el orden de las filas o de las columnas, sale otra OR, ya que hay otra confluencia de modalidades en la casilla a_1. Si los datos son apareados, la casilla de referencia es la a_2 , comparada con la b_1
La <u>hipótesis nula, H_0</u> , presupone que la OR vale 1 . Pero la OR sola es un valor puntual y no sirve para la valoración estadística; hay que calcular el intervalo de confianza, que veremos enseguida. **Si el intervalo no incluye el 1, se rechaza la hipótesis nula** y se concluye que hay una asociación significativa al nivel de significación que hayamos elegido para c ó t y en el sentido que indique la casilla de referencia. <u>Si el intervalo incluye el 1, no puede rechazarse H_0</u>

Cálculo del intervalo de confianza de una OR

El método más sencillo utiliza logaritmos. Se halla el logaritmo neperiano de la OR y a éste se le suma y resta el error muestral E, que tiene una fórmula fácil (habitualmente se toma un nivel de significación alfa para c ó t de 0,05).
Así tendremos los límites del intervalo, cuyos antilogaritmos son los límites del IC de la OR

a) **DATOS INDEPENDIENTES**

$$\textbf{IC del ln OR} = \ln OR \pm c\sqrt{\frac{1}{a_1}+\frac{1}{a_2}+\frac{1}{b_1}+\frac{1}{b_2}}$$ **; si N<30, en vez de c se toma t_{N-2}**

hallado el intervalo se calculan los antilogaritmos (e^x) de los límites del intervalo: son los límites del IC de OR

Ejemplo: Se estudia en una comarca la mortalidad precoz (antes de los 60 años) en fumadores y no fumadores.

		Fumador		
		Si	No	
Muerte precoz	Si	700	200	900
	No	300	300	600
		1000	500	1500

La **OR vale 3,5** ; la probabilidad de muerte precoz de un fumador es 3,5 veces mayor que la un no fumador.
El ln de la OR es 1,252762968 (hay que seguir trabajando al menos con 6 decimales)

IC del ln OR = 1,252763 ± 1,96 $\sqrt{1/700 + 1/200 + 1/300 + 1/300}$ =
1.252763 ± 0,224291 = ∈(1,028472÷1,477054)
Calculando los antilogaritmos de ambos límites (y redondeando a dos decimales) :
IC de OR = € (2,80 ÷ 4,38) , que es significativo al no estar el 1 en el intervalo. nivel de significación 0,05 (hemos tomado para c el valor de 1,96). Asociación positiva entre fumar y muerte precoz.

b) **DATOS APAREADOS**

La fórmula es la misma, excepto lo que va dentro de la raíz cuadrada: $\sqrt{\frac{1}{a_2}+\frac{1}{b_1}}$

Sólo se tienen en cuenta los datos discordantes. Al ser datos apareados N = a2 + b1.
La OR va referida a la casilla a_2 (comparada con la b_1)

Ejemplo: Se comparan en 62 pacientes la eficiencia de dos análisis distintos (A y B) en el diagnóstico de una enfermedad.

		A		
		+	-	
B	+	**20**	**12**	32
	-	**15**	**15**	30
		35	27	62

OR = 0,8 ; ln 0,8 = -0,223144 ; **N** = 12 + 15 = 27 (los discordantes!) IC lnOR = -0,223144 ± 2,060 √ 1/12 + 1/15 = -0,223144 ± 0,797835 = € (-1,020979 ÷ 0,574691) . Sus antilogaritmos son los límites de OR (redondeamos a dos decimales) : **IC de OR = € (0,36 ÷ 1,78)** La OR no es significativa al incluir al 1 en su intervalo. Es n.s. p>0,05 . Ambos análisis son igual de eficientes, aunque B parezca algo inferior, ya que la OR de 0,8 indica según las casillas a_2 y b1 que es inferior en acertar cuando el otra análisis falla.

Riesgo relativo (RR)

Es el parámetro típico de los estudios de cohortes, que son estudios prospectivos en los que se siguen durante años a personas expuestas y no expuestas a un determinado riesgo o condición para ver si enferman o no. Por ejemplo, el seguimiento durante años de personas que toman un determinado medicamento para prevenir enfermedades graves y de un grupo control que no lo toma. En vez de medicamentos el objeto de estudio puede ser el ejercicio físico u otros hábitos saludables, psicoterapia, etc. Aunque se habla de riesgo, a veces se trata de un beneficio. Problemas del lenguaje.
Matemáticamente es siempre posible calcular el RR, con independencia de que sea un estudio caso-control o de cohortes. Pero cada uno tiene su parámetro adecuado. Si el riesgo es escaso (< 0,1 ó 10%) OR y RR toman valores muy parecidos, pero a medida que el evento se hace más frecuente empiezan a separarse cada vez más. En muchos estudios se usa la OR como equivalente del RR, lo que no es correcto.

El RR es el cociente de los riesgos de expuestos y no expuestos.. Se expresa como proporción o porcentaje.

Se parte de la tabla de 2x2 :

		Enfermedad o evento negativo		
		+	-	
Exposición o factor a estudio	+	$\mathbf{a_1}$	$\mathbf{a_2}$	Na
	-	$\mathbf{b_1}$	$\mathbf{b_2}$	Nb
		N_1	N_2	N

$$RR = \frac{a_1}{N_a} : \frac{b_1}{N_b} = \frac{a_1 N_b}{b_1 N_a}$$

La hipótesis nula H_0 es que RR = 1. **La valoración es similar a la de la OR**. Para ver si la asociación es significativa, es preciso calcular el intervalo de confianza de RR.

El RR es significativo si su IC no incluye al 1

Cálculo del intervalo de confianza de RR

--Se calcula el IC del logaritmo neperiano del RR y luego se vuelve a "números normales"... Así:

$$\textbf{IC del ln de RR} = \ln R \pm c\sqrt{\frac{1}{a_1}+\frac{1}{b_1}-\frac{1}{N_a}-\frac{1}{N_b}}$$

¡ojo a los dos signos menos!
si N es menor de 30, en vez de **c** se toma **t** con gl N-2

--luego se calculan los antilogaritmos (e^x) de los extremos del intervalo :
son los límites del IC del RR **IC = ∈ (límite inferior ÷ límite superior)**

Ejemplo:
En un hospital inglés se aplicó un programa destinado a incrementar la duración de la lactancia materna. A los 3 meses ya no daban el pecho 32 de las 51 mujeres del grupo de intervención y 52 de las 57 del grupo control. Concluyen que con el programa han reducido claramente el riesgo de abandono de la lactancia materna a los 3 meses.
Veamos:

Programa fomento Lactancia Materna (LM)

		Abandono		
		+	-	
Programa	+	32	19	51
	-	52	5	57
		84	24	108

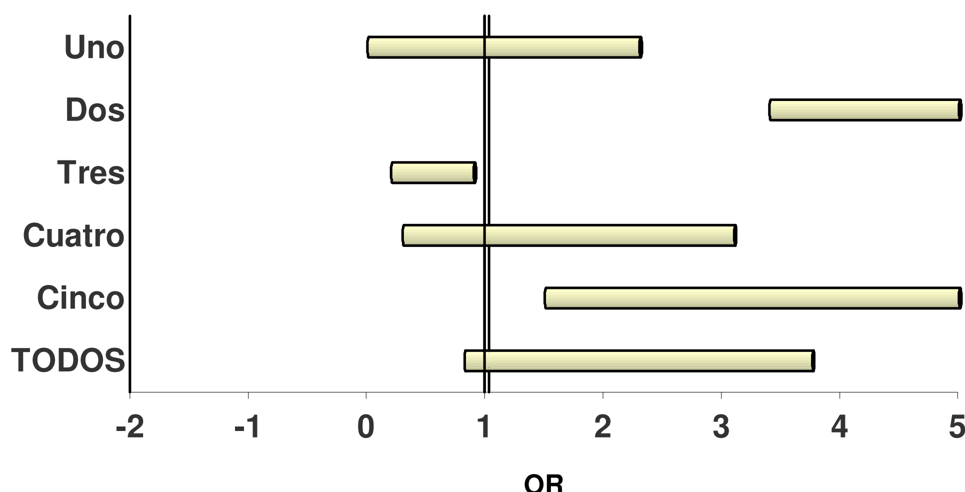
RR de abandono de la LM en las que han seguido el programa:

RR = (32 * 57)/(52 * 51) = **0,688**

Al ser la RR < 1 indica que el riesgo es menor para la situación que indica la casilla a1, es decir, abandono habiendo seguido el programa. Pero este riesgo menor ¿es significativo? . Para contestar a esta pregunta hay que calcular el IC de RR, que aplicando la fórmula resulta ser $\in$ **(0,55 ÷ 0,86)** , que al no incluir el 1 es significativo al nivel de significación empleado, que es 0,05 (ya que se ha tomado c = 1,96)

Metaanálisis

Con frecuencia se observa que estudios sobre un mismo tema dan resultados divergentes, incluso con grandes diferencias. En estos casos es de ayuda la técnica llamada Metaanálisis, que permite calcular un IC conjunto para todos los estudios y de él sacar la conclusión adecuada. Es un procedimiento muy complejo y laborioso, en el que no entramos (está muy bien descrito en el libro de Armitage/Berry). Como orientación se pueden hacer dos cosas: 1) pasar a un gráfico los IC de las diversas OR, lo que nos da una idea del conjunto 2) a partir de una tabla que englobe el total de los datos de todos los estudios, calcular la OR y su IC por el procedimiento ordinario en vez de por el complicado método ortodoxo.
El siguiente gráfico representa gráficamente un metaanálisis:



OR . Aclaraciones sobre la tabla y repaso de la valoración

Con los datos que se dan en el enunciado de un problema, la tabla se puede construir de 4 formas distintas, que nos dan dos OR diferentes, pero relacionadas. Cada OR es la inversa de la otra (1/OR). En los límites de confianza el inverso del Li de una OR es el Ls de la otra y viceversa.

Ejemplo

Pinto y col. han estudiado en una zona de México la relación entre malformaciones congénitas y consanguinidad en 33194 recién nacidos en un periodo de 6 años. Hubo 1117 neonatos con alguna anomalía congénita. Se tomó como control de cada caso al primer neonato sano del mismo sexo nacido después. 21 de los malformados tenían el antecedente de consanguinidad por 8 de los controles. Valore el resultado (por OR).

Se trata de un contraste de dos variables cualitativas con dos modalidades cada una : Malformación (Sí , No) y Consanguinidad (Sí , No) . Los datos son independientes. La hipótesis nula H_0 es que no hay diferencias significativas en las malformaciones que aparecen en niños con y sin antecedente de consanguinidad,o sea una OR =1 . Este problema se puede resolver por la fórmula nº 2 ó 1, pero se pide que se haga valorando la OR.

Pasos:

1---construir la tabla de 2x2 ; ocurre que podemos construir 4 tablas distintas. Calcularemos en cada una la OR y su IC (se ha tomado una c = 1,96 que corresponde a $\alpha = 0{,}05$))

1		Malformaciones		
		Sí	No	
Consanguinidad	Sí	21	8	29
	No	1096	1109	2205
		1117	1117	2234

OR = 2,66 (2,65613...)

$\in$ (1,171÷6,022) (1,171487... y 6,022306...)

2		Malformaciones		
		No	Sí	
Consanguinidad	Sí	8	21	29
	No	1109	1096	2205
		1117	1117	2234

OR = 0,38 (0,376486...)

$\in$ (0,166÷0,854) (0,166049... y 0,853615...)

3		Malformaciones		
		Sí	No	
Consanguinidad	No	1096	1109	2205
	Sí	21	8	29
		1117	1117	2234

OR = 0,38

$\in$ (0,166÷0,854)

4		Malformaciones		
		No	Sí	
Consanguinidad	No	1109	1096	2205
	Sí	8	21	29
		1117	1117	2234

OR = 2,66

$\in$ (1,171÷6,022)

Se obtiene pues dos OR distintas.

El nº inverso de la primera OR es 1/2,65613 ≈ 0,38 (la otra OR) y el inverso de la segunda OR es 1/0,3746486 ≈ 2,66

Y para el intervalo de confianza : 1/1,171487 ≈ 0,854 y 1/6,022306 ≈ 0,166
1/0,166049 ≈ 6,022 y 1/0,853615 ≈ 1,171

La valoración de la OR se hace por la casilla a_1 .
Recuerden la nomenclatura de las casillas:

	a_1	a_2	Na
	b_1	b_2	N_b
	N_1	N_2	N

En la tabla 1 la casilla a_1 es la confluencia de malformación y consanguinidad; como la OR (2,66) es >1 , interpretamos que cuando hay consanguinidad, se observan más malformaciones de lo esperado. Esta asociación es estadísticamente significativa al no estar el uno en el intervalo de confianza ($p<0,05$). La tabla 4 es lo mismo, pero visto desde el lado opuesto. Se asocian no malformación y no consanguinidad.
En la tabla 2 la casilla a_1 corresponde a consanguíneos no malformados; su OR = 0,38 , que es <1 , es decir que los niños consanguíneos sin malformación son menos de los esperados y además de forma significativa ($p<0,05$) al no incluir el 1 su intervalo de confianza. En la tabla 3 confluyen malformación y no consaguinidad, con valoración similar.

¿Cuál elegir?
La que mejor se corresponda al objetivo del problema, que en este caso es valorar una posible asociación entre consanguinidad y malformaciones congénitas. Por tanto la mejor tabla es la nº 1, que lo hace de forma directa, seguida de la 2 . Pero todas son buenas y nos llevarán a la misma conclusión, aunque por caminos más retorcidos y menos intuitivos.

Un razonamiento similar se puede hacer para el RR

Puntos débiles de las OR

La OR es otra forma de enfocar el contraste de frecuencias de dos variables cualitativas con dos modalidades cada una. La decisión estadística es la misma.
Es un parámetro que se puso de moda en el pasado decenio. Es muy útil, pero tiene también sus puntos débiles, **los mismos que el procedimiento clásico**. Recordémoslos:
--las muestras de casos y controles con frecuencia no son aleatorias. Siempre hay que preguntarse si todos los individuos de las poblaciones de casos y controles han tenido la misma probabilidad de salir elegidos para el estudio.
--Los criterios de exclusión del estudio a veces no son los mismos para casos y controles.
--Hay que vigilar los sesgos de recuerdo ("recall bias") en la documentación clínica, pues los pacientes son reiteradamente preguntados sobre los factores de riesgo, cosa que no les ha sucedido a los controles.
--Hay que buscar la posible existencia de factores de confusión, que pueden simular asociación significativa entre exposición y enfermedad. Por ejemplo, un estudio puede sugerir que los alcohólicos tienen un riesgo elevado de padecer cáncer de pulmón, hasta que se descubre que prácticamente todos los alcohólicos eran fumadores. Otro ejemplo: en muchas ocasiones se prescriben estrógenos para las hemorragias vaginales. Si meses después se descubre un cáncer de útero, podría pensarse que es un efecto secundario de los estrógenos. Pero no hay que olvidar que las hemorragias son un síntoma de cáncer uterino.

Si se identifican “confundidores” hay que estratificar en subgrupos del confundidor. Los más frecuentes son: edad, sexo, nivel sociocultural, tabaco, alcohol, drogas....
--No se debe olvidar que una relación o asociación significativa sólo permite concluir causalidad si el estudio es experimental.

En los ejercicios que hemos realizado por los contrastes clásicos, se puede también calcular la OR, aunque no sea el parámetro más adecuado. Pero se llega a las mismas conclusiones:

Ejercicio	Variables	Datos	a1	a2	b1	b2	OR	IC- OR	¿rechazo de H_0?
1.1 y 1.2	Fumar (sí , no) Sexo (♂ , ♀)	Independientes	70	30	80	120	3’50	2’10 5’84	SI Hombres fuman más
1.3	Ciudad (X , Y) Colesterol (alto, bajo)	Independientes	6	14	15	15	0’43	0’13 1’42	NO
1.4	Miopía (si-no) Sexo (♂ - ♀)	Independientes	4	16	7	21	0,75	0’19 3’01	NO
1.5	BCG (si-no) TBC (sí , no)	Independientes	10	80	60	40	0’08	0’04 0’19	SI Si BCG, menos TBC
2.1	Análisis (A-B) Result. (+ -)	Apareados	15	5	15	65	0’33	0’11 0’99	SI A es mejor
3	Radar (X , Y) Aviso (sí , no)	Apareados	20	5	3	5	1’67	0’28 9’95	NO

La OR se debe reservar para los estudios caso-control, aunque siempre es calculable.

Tema 17 : CONTRASTE DE UNA VARIABLE CUALITATIVA Y OTRA CUANTITATIVA

Se concreta en un contraste de 2 ó más medias. Los datos pueden ser independientes, en los que los problemas de comparación y relación se resuelven por las mismas fórmulas, o apareados, en cuyo caso hay que distinguir muy bien si se trata de una comparación o de una relación, ya que las fórmulas a utilizar son distintas.
Hay que plantearse la pregunta: ¿Me piden que busque si hay diferencias entre los grupos o muestras contrastados o bien si hay una relación, una dependencia entre ellos?. En la tabla guía del tema 15 pueden verse las diversas situaciones que se nos pueden plantear y la forma de abordarlas.

1) La variable cualitativa tiene dos modalidades y los datos son independientes.

Se trata de un contraste de dos medias. Para resolverlos se dispone de una prueba paramétrica ,que llamamos fórmula nº 6, y de otra no paramétrica, la prueba de Mann-Whitney

Fórmula nº 6

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}} \text{ , siendo } s^2 = \frac{s_1^2(N_1 - 1) + s_2^2(N_2 - 1)}{N_1 + N_2 - 2}$$

s^2 es la varianza común

H_0: no hay diferencias significativas entre las medias contrastadas ; las diferencias numéricas observadas se explican por el azar.

Condición de aplicación para muestras pequeñas : que el cociente da varianzas, V, obtenido al dividir la varianza mayor por la menor, no supere el valor de referencia de F. Con independencia del orden con que nos den los datos, la muestra nº 1 será la de varianza mayor y la de varianza menor será la nº 2. $V < F(N_1\text{-}1 ; N_2\text{-}1; 0'05)$ .Si no cumple la condición, hay que pasar de oficio a la prueba no paramétrica.

Valoración: si ambas muestras son grandes por la DN ; si alguna es pequeña por $t(N_1 + N_2\text{-}2 ; \propto)$
Si Z < valor de referencia : no puede rechazarse H_0 , no se han encontrado diferencias significativas. (suele escribirse :N.S. ó n.s.)
Si Z ≥ valor de referencia ; se rechaza H_0 al nivel de significación probado (y suele escribirse $p < 0{,}05$ ó $p < 0{,}01$ ó $p < 0{,}001$) y se acepta la hipótesis alternativa, H_1 Hay que dar el sentido.
Recuerdo que, de no decirse lo contrario, si se supera un nivel de siginificación, hay que probar con el siguiente...

Ejercicio 17-1
Se mide la talla en muestras de adultos jóvenes de los pueblos A y B
En A obtenemos: $x_1 = 169$ cm , $s_1 = 5$ cm , N=100
En B : $x_2 = 166$ cm , $s_2^2 = 16\ cm^2$, N = 80.
¿Puede afirmarse que los de son más altos que los de B?

**Se trata de una prueba de comparación entre una variable CL, PUEBLO, con dos modalidades, A y B, y otra variable CT, TALLA, medida en los individuos de las muestras de A y B. Los datos son independientes. Contraste de dos medias. A resolver en principio por la fórmula nº 6.
** H_0 : no hay diferencias significativas entre las tallas de A y B = los de A no son más altos que los de B
**Las muestras son grandes y no hay condición de aplicación que comprobar.

**Se calcula la varianza común s^2

$$s^2 = \frac{(25 * 99) + (16 * 79)}{100 + 80 - 2} = 21$$

**Se calcula Z

$$Z = \frac{169 - 166}{\sqrt{\frac{21}{100} + \frac{21}{80}}} = 4,36$$

** Valoración: por los valores de c de la DN correspondientes a los niveles de significación habituales

Z = 4'36 es > que c0'05 = 1'96 y también a c0'01 = 2'58 y a c0'001 = 3,30

Por tanto se rechaza H_0 al nivel de significación de 0'001 y se acepta H_1: las tallas no son iguales; hay diferencias significativas entre ellas. Sentido : la media de A es más alta que la de B.

**Y contestando a la pregunta que nos han hecho : Sí

En un examen hay que seguir fielmente los pasos del ejercicio anterior En los siguientes, por ahorro de espacio, se hará de forma más telegráfica

Ejercicio 17-2

En 15 soldados se mide la concentración de la proteína P en la sangre (en mg/dl). En 5, oriundos de la provincia A, obtenemos lo siguiente: 5 , 7 , 6 , 7 , 5. En los 10 restantes, que proceden de la provincia B: 8 , 10 , 11 , 8 , 8 , 7 , 7 , 6 , 7 , 8 . ¿Hay diferencias entre ambas provinicas? ¿Puede decirse que las diferencias se deben a la excelente calidad del agua de B?

Contraste de una Vble. CL , Provincia, con 2 modalidades, A y B, y otra CT, concentración sanguínea de P. Datos independientes.

→ fórmula nº 6 . Al ser muestras pequeñas hay que comprobar si cumple la condición de aplicación.

H_0 : no hay diferencias significativas entre A y B

Como nos dan los datos originales, hay que calcular la media y la varianza de cada grupo.

	Media	Varianza	N
A	6	1	5
B	8	2'22	10

Como la varianza de B es mayor que la de A, la muestra 1 será B y la 2 será A

V=2'22/1=2'22 que es < F(9 ; 4 ; 0'05)= 6'00 y por tanto cumple la condición y podemos seguir

s^2 = 1'84 y

$$Z = \frac{6 - 8}{\sqrt{\frac{1,84}{5} + \frac{1,84}{5}}} = -2,69$$

|Z| > t(13 ; 0'05) = 2,160, por lo que se rechaza H_0 al nivel de significación de 0'05: hay diferencias entre los soldados de A y B ; sentido: los soldados de B tienen la proteína P significativamente más alta.

Y contestando a la otra pregunta: no lo podemos saber....

Ejercicio 17-2 bis

Resolver el problema anterior por una prueba no paramétrica.

La prueba es la **nº 7 , de Mann-Whitney**.
Como prueba no paramétrica no tiene condiciones de aplicación.
Consiste en ordenar a todos los individuos en conjunto, asignándoles un nº de orden. La ordenación se puede hacer de mayor a menor o de menor a mayor. En caso de empate a cada individuo se le asigna la media de los números de orden que habría que repartir entre ellos.
El nº de orden que se adjudica a cada dato se anota en la columna de R que le corresponde. A la suma de las columnas de R, las llamamos, respectivamente, R_1 y R_2

Para aplicar la fórmula se toma para R el valor de la menor de R_1 y R_2 , con su n correspondiente.

$N = n_1 + n_2$.
Se valora por la DN (si $N \geq 30$) ó por $t(N-2 , \alpha)$
Una forma práctica de resolverlo es utilizar una plantilla como la que se ofrece a continuación

Se ordenan todos los datos a la vez

Individ.		X_1	R		X_2	R
1						
2						
3						
4						
5						
6						
.....						
n						
		Suma			Suma	
			R_1			R_2

$N = n_1 + n_2$

$$Z = \frac{R - \frac{n(N+1)}{2}}{\sqrt{\frac{n_1 n_2 (N+1)}{12}}}$$

R es la menor de R1 y R2 ; **n** es el tamaño de la muestra que corresponde a esa R

Prueba de que se han calculado bien las R : **$R_1 + R_2 = N(N+1)/2$**

Si hay diferencias significativas, hay que dar el sentido: la media más alta es la del grupo con R mayor (si hemos ordenado de menor a mayor)

****en el problema propuesto:
Es un contraste de una Vble. CL , Provincia, con 2 modalidades, A y B, y otra CT, concentración sanguínea de P. Datos independientes
H_0 : no hay diferencias significativas entre A y B

Individ.		$X_1 = B$	R		$X_2 = A$	R
1		8	*11'5*		5	*1'5*
2		10	*14*		7	*7*
3		11	*15*		6	*3'5*
4		8	*11'5*		7	*7*
5		8	*11'5*		5	*1'5*
6		7	*7*			
7		7	*7*			
8		6	*3'5*			
9		7	*7*			
10		8	*11'5*			
		Suma	*99'5*		**Suma**	*20'5*
			R_1			R_2

$N = n_1 + n_2 = 15$

$$Z = \frac{20{,}5 - \frac{5*16}{2}}{\sqrt{\frac{10*5*16}{12}}} = -2{,}388$$

Prueba:
R1+R2: 99,5+20,5 = 120
N(N+1)/2: 15*16/2 = 120

Se valora por t con g.l. de 13 : $|Z| > t(13, 0'05) = 2'160$
Por tanto se rechaza H_0 al nivel de significación de 0'05 y se acepta H_1 : sí hay diferencias , $p<0{,}05$

La prueba no paramétrica, aunque menos potente, también ha logrado descubrir las diferencias
Las preguntas se responden como en el ejercicio anterior

Nota:
** Hay un procedimiento clásico de resolver el Mann-Whitney. Se calculan dos posibles resultados:
$Z_1 = n_1 n_2 + n_1 (n_1+1)/2 - R_1$ y $Z_2 = n_1 n_2 + n_2 (n_2+1)/2 - R_2$
Se toma como resultado final, Z, el menor de los dos y se compara con un valor de referencia en una tabla especial, la tabla de la U, para tomar la decisión estadística. no vemos aquí este método.
** La fórmula que utilizamos, la nº 7, es válida a partir de un tamaño muestral pequeño, que algunos cifran en 5, y tiene la ventajas obre el procedimiento clásico de poder ser valorada por la DN o la t de Student.
** Hay una variante de nuestra fórmula 7, que tiene en cuenta el menor de Z_1 y Z_2, y se valora también por la DN o la t . Sólo cambia el numerador, que es : Zmenor - $(n_1 n_2 / 2)$

---

2) La variable cualitativa tiene más de dos modalidades con datos independientes

Es un contraste de 3 o más medias, cuyo método paramétrico es el análisis de la varianza, más conocido como **ANOVA** (de su nombre en inglés: ANalyisis Of Variance). Hay varios ANOVAs ; aquí utilizaremos el **ANOVA-1** (también conocido como One Way ANOVA). Se analiza un factor (más adelante se verá el ANOVA-2, que analiza dos factores) utilizando las varianzas.
Se necesitan los datos originales para el cálculo clásico, que es bastante farragoso y que se facilita utilizando la plantilla siguiente :

Muestras →	1		2		3		……		k		
Individuos ↓	**X**	**X^2**	**X**	**X^2**	**X**	**X^2**	**X**	**X^2**	**X**	**X^2**	**Valoración**: **por F(k-1 , N-k , α)**
1											
2											
3											
4											
5											
6											
………											**Si se rechaza H_0 hay que aplicar prueba de Scheffé, de dos en dos, ordenados por su media**
ΣX											ΣΣX = B
$(\Sigma X)^2$											
n											**Σn = N**
$(\Sigma X)^2/n$											**$\Sigma[(\Sigma X)^2/n] = A$**
ΣX^2											**$\Sigma(\Sigma X^2) = C$**
$\bar{X}$											

$$C_A = A - \frac{B^2}{N} = \square \qquad V_A = \frac{C_A}{k-1} = \square$$

$$C_T = C - \frac{B^2}{N} = \square \qquad Z = \frac{V_A}{V_R} = \square$$

$$V_R = \frac{C_R}{N-k} = \square$$

$$C_R = C_T - C_A = \square$$

En la mayoría de los programas estadísticos se utiliza una nomenclatura distinta a la usada aquí :

C_A es llamada variación inter ó entre grupos ("between") , la que procede del objeto de estudio
C_R es llamada variación intra o variación residual (·within") , la que procede de los individuos
C_T es la variación total, suma de las otras dos

Los números suelen ir bajo el epígrafe "suma de cuadrados" o "ssq" ´o "msq"

V_A es la varianza inter ; V_R es la varianza intra En vez de Z ponen F

¿Y si no se conocen los datos originales?
Conociendo la media, la varianza y el tamaño de cada uno de los grupos se pueden calcular sus respectivos ΣX y ΣX_2, por las fórmulas siguientes, que están el página 15 del Formulario:

$$\sum \mathbf{X} = \mathbf{n}\bar{\mathbf{X}}$$

$$\sum \mathbf{X}^2 = \frac{\mathbf{s}^2\mathbf{n}(\mathbf{n}-1) + \left(\sum \mathbf{X}\right)^2}{\mathbf{n}}$$

y pueden colocarse en su sitio en la plantilla anterior

El ANOVA-1 es una prueba muy robusta y no es preciso comprobar condiciones de aplicación.
Si la prueba lleva a rechazar H_0 , la conclusión es que los grupos, las k medias, difieren entre sí. Pero ésto no quiere decir que estas diferencias existan en todos los casos cuando las tomamos dos a dos. Puede ocurrir que sólo alguna o algunas de las medias sean las responsables de las diferencias. Para averiguar ésto se dispone de varios métodos. El aquí elegido es el **método de Scheffé**, cuya metódica se verá más adelante.

Ejercicio 17-3

A 4 grupos de cobayas se les alimenta con dietas distintas (cada grupo dieta distinta). Al cabo de unos días se comprueba su ganacia de peso en gramos :

Dieta A : 32 , 37 , 34 , 30 , 33
Dieta B : 36 , 38 , 37 , 30 , 34 , 39
Dieta C : 35 , 30 , 36 , 29 , 31 , 29
Dieta D : 29 , 31 , 39 , 39 , 28
Valorar el resultado

*Problema de contraste entre una variable CL, DIETA, con 4 modalidades, A – B – C – D, y otra CT, ganacia de peso.
Los datos son independientes. Por tanto es un contraste de k medias, a resolver por ANOVA-1
H_0 : no hay diferencias significativas entre las medias de los grupos contrastados; las variaciones de las medias se deben al azar
Para los cálculos utilizaremos la plantilla correspondiente

Muestras →	**1 A**		**2 B**		**3 C**		**4 D**		
Individuos ↓	**X**	**X^2**	**X**	**X^2**	**X**	**X^2**	**X**	**X^2**	**Valoración**: por **F (k-1 , N-k , α)**
1	32	1024	36	1296	35	1225	29	841	
2	37	1369	38	1444	30	900	31	961	
3	34	1156	37	1369	36	1296	30	900	
4	30	900	30	900	29	841	30	900	
5	33	1089	34	1156	31	961	28	784	
6			39	1521	29	841			
………									
									Si se rechaza H_0 hay que aplicar prueba de Scheffé, de dos en dos, ordenados por su media
ΣX	166		214		190		148		ΣΣX = B **718**
$(ΣX)^2$	27556		45796		36100		21904		
n	5		6		6		5		**Σn = N 22**
$(ΣX)^2/n$	5511'2		7632'67		6016'67		4380'6		**$Σ[(ΣX)^2/n]$ = A 23541'33**
$ΣX^2$		5538		7686		6064		4386	**$Σ(ΣX^2)$= C 23674**
$\bar{X}$	**33'20**		**35'67**		**31'67**		**29'60**		

$C_A = A - \frac{B^2}{N}$ = 108’4 $V_A = \frac{C_A}{k-1}$ = 36’14

$C_T = C - \frac{B^2}{N}$ = 241’09 $Z = \frac{V_A}{V_R}$ = **4’90**

$V_R = \frac{C_R}{N-k}$ = 7’37

$C_R = C_T - C_A$ = 132’67

Valoración: por F(3 ; 18 ; ∝) : para 0’05 vale 3’16 y para 0’01 vale 5’09; $Z > F_{0'05}$ y por tanto se rechaza H_0 y se acepta H_1 : hay diferencias significativas entre el conjunto de las medias contrastadas. Esto nos obliga a realizar la prueba de Scheffé, fórmula 8 bis

Prueba de Scheffé

Pasos

1) ordenar las medias, de mayor a menor o viceversa
2 compararlas por parejas, empezando por las màs dispares, las de los extremos
3) aplicar la fórmula 8 bis

$$Z_{sch} = \frac{(\overline{X}_i - \overline{X}_j)^2}{V_R(k-1)(\frac{1}{n_i} + \frac{1}{n_j})}$$

Valoración por $F_{k-1, N-k}$

Los datos los tomamos del cálculo del ANOVA-1 . En el numerador están las medias de los dos grupos. En el denominador aparte de V_R están el nº de grupos o muestras (k) y los tamaños de las dos muestras que estamos contrastando (n_i y n_j).

4) la Z obtenida se contrasta con la F de referencia y se toma la decisión estadística

En el problema que nos ocupa:

El orden es:					
	muestra →	B	A	C	D
	Media →	35’67	33’20	31’67	29’60

*** comparamos B y D

$$Z = \frac{(35'67 - 39'60)^2}{7'37 * 3 * (\frac{1}{6} + \frac{1}{5})} = 4'54$$

Contrastamos Z con F . Es mayor que la F(3 ; 18 ; 0’05)=3’16 y por tanto se rechaza la hipótesis nula y se acepta la alternativa en el sentido de que **B es superior a D**.

*** esto obliga a seguir probando, ahora con B y C

$$Z = \frac{(35'67 - 31'67)^2}{7'37 * 3 * (\frac{1}{6} + \frac{1}{6})} = 2'17$$

Aquí Z es menor que la F de referencia y por tanto **no hay rechazo de H_0**

*** no hace falta probar con B y A, pues nos darán una Z aún más baja

*** sí que hay que probar A y D

$$Z = \frac{(33'20 - 29'60)^2}{7'37 * 3 * (\frac{1}{5} + \frac{1}{5})} = 1'47$$

Z también es menor que la F de referencia y por tanto no hay rechazo de H_0

*** no hace falta seguir probando, ya que las Z que obtengamos serán aún menores.

Conclusión final : La prueba de ANOVA-1 nos dice que las ganancias de peso conseguidas con las cuatro dietas son significativamente distintas en su conjunto. La prueba de Scheffé nos aclara que ello se debe fundamentalmente a la superioridad de B sobre D.

PRUEBA DE KRUSKAL-WALLIS

Como el ANOVA-1 es una prueba muy robusta y no comprobamos condiciones de aplicación, no se nos remite de oficio a la prueba no paramétrica correspondiente, que es la de Kruskal-Wallis.

Esta prueba al ser no paramétrica no tiene condiciones de aplicación. Funciona de forma similar al Mann-Whitney. Se ordenan todos los individuos en conjunto, asignándoles un nº de orden. La ordenación se puede hacer de mayor a menor o de menor a mayor. En caso de empate a cada individuo se le asigna la media de los números de orden que habría que repartir entre ellos.

Una forma práctica de resolverlo es utilizar una plantilla como la que se ofrece en el Formulario y que vemos ahora para resolver el problema anterior por la prueba de Kruskal-Wallis.

Ejercicio 17-3 bis

Resolver el ejercicio anterior por una prueba no paramétrica.

Para las variables de este supuesto la prueba adecuada es la de Kruskal-Wallis

Individ.	**Muestras**									
	1 A		**2 B**		**3 C**		**4 D**		***** Se ordenan los datos de todas las muestras en conjunto. *** Valoración por χ2 (k-1 , α) *** Si se rechaza H_0 hay que aplicar la prueba de Mann-Whitney de dos en dos, ordenados por su T**	
	x_1	**R**	**x_2**	**R**	**x_3**	**R**	**x.**	**R**		
1	32	*12*	36	*17'5*	35	*16*	29	*3*		
2	37	*14'5*	38	*21*	30	*7*	31	*10'5*		
3	34	*14'5*	37	*19'5*	36	*17'5*	30	*7*		
4	30	*7*	30	*7*	29	*3*	30	*7*		
5	33	*13*	34	*14'5*	31	*10'5*	28	*1*		
6			39	*22*	29	*3*				
										Suma
T = ΣR	66		101'5		57		28'5			
T^2	4356		10302'25		3249		812'25			
n	5		6		6		5		**→ N = Σn →**	**22**
T^2/n	871'2		117'04		541'5		162'25		**→ Σ(T^2/n) →**	**3292'19**

$$Z = \left[\frac{12}{N(N+1)} \sum \left(\frac{T^2}{n} \right) \right] - 3(N+1)$$

Z > χ2(3 ; 0'05)=7'81 y por tanto se rechaza H_0 a ese nivel de significación y la conclusión es que los grupos en conjunto difieren significativamente. Para saber que grupos son los que más contribuyen a estas diferencias se aplica la prueba de Mann-Whitney de dos en dos, ordenados por su T

El orden es B -A–C–D. Se empieza comparando los grupos más dispares y se sigue así en orden decreciente.

Resumiendo:

	R_1	R_2	n_1	n_2	Z	t(N-2 ; 0'05)	¿significativo?
B y D	49	17	6	5	-2'37	2'262	si
B y C	49	29	6	6	-1'60	2'228	no
...							no
A y D	38	17	5	5	-2'19	2'306	no

La conclusión es la misma que en la prueba de Scheffé: las diferencias se deben fundamentalmente a la superioridad de B sobre D

3) La variable cualitativa tiene dos modalidades y los datos son apareados.

Se trata de un contraste de dos medias. Al ser los datos apareados hay que distinguir muy bien si es un problema de **comparación**, en cuyo caso se toman las fórmulas 10 u 11, o bien si es un problema de **relación**, a resolver por las fórmulas 14 ó 15

3-a : problema de comparación

Primero hay que calcular las diferencias entre los pares de valores y luego calcular la media y la varianza de estas diferencias (para la varianza necesitamos también los cuadrados de las diferencias). Con ello ya se puede aplicar la fórmula nº 10

$$Z = \bar{X}_d \sqrt{\frac{N}{s_d^2}}$$ **Valoración: muestra grande por DN; si pequeña por t_{N-1}**

H0 : no hay diferencias entre los datos comparados

Es útil disponerse una tabla auxiliar cuyos encabezados sean: X Y x_d x^2_d

Ejercicio 17-4

Probamos el efecto de un somnífero en 15 personas midiendo las horas que duermen tomándolo y sin tomarlo.

Individuos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
horas CON	12	5	13	10	13	10	8	8	7	6	9	8	7	7	5	
horas SIN	8	6	8	6	10	9	4	7	6	6	8	6	9	7	6	¿Es efectivo?

Solución: Problema de contraste de una variable cualitativa , TIPO DE SUEÑO (CON, SIN) y otra cuantitativa, HORAS DE SUEÑO . Datos apareados. Es un problema de comparación, a resolver por la fórmula nº 10. H0 : no hay diferencia entre las horas dormidas en ambas situaciones

Indiv	X	Y	x_d	x^2_d
1	12	8	4	16
2	5	6	-1	1
3	13	8	5	25
4	10	6	4	16
5	13	10	3	9
6	10	9	1	1
7	8	4	4	16
8	8	7	1	1
9	7	6	1	1
10	6	6	0	0
11	9	8	1	1
12	8	6	2	4
13	7	9	-2	4
14	7	7	0	0
15	5	8	-1	1
Suma			22	96

$\Sigma d = 22$, $\Sigma d^2 = 96$, $\bar{X} = 1,47$, $s^2 = 4,552$

$z = 1,47 \sqrt{15 / 4,552} = 2,67$ que es mayor que $t(14, 0'05) = 2,145$

Se rechaza H_0 a ese nivel de significación y se acepta H1: hay diferencias significativas entre las horas dormidas tomando y sin tomar el medicamento.

Sentido: tomándolo se duerme más

.Ejercicio 17-4 bis
----Resuelva el ejercicio anterior con una prueba no paramétrica

La prueba no paramétrica es el **test de los signos**.
Se compara el par de valores de cada individuo y se anota un signo (+ , - , 0) según el criterio que se adopte: por ejemplo, "+" si el primer dato es mayor , "-" si es menor y "0" si son iguales. También puede hacerse todo lo contrario, ya que el resultado no variará, pues se toma siempre el signo mayoritario.
Se cuentan los signos "+" y "-". Uno de ellos, cualquiera, se asigna a N_1 y el otro a N_2 . $N=N_1+N_2$
Para la fórmula se toma la la mayor de N_1 y N_2 y para evitar confusiones con las" enes" la llamamos x.

Fórmula nº 11 (Test de los signos)

$$Z = \frac{(2x - N)}{\sqrt{N}}$$

siendo x el mayor de N_1 y N_2

valorar **por t_{N-1} ó DN (si N ≥ 30)**

	Individuos														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
horas CON	12	5	13	10	13	10	8	8	7	6	9	8	7	7	5
horas SIN	8	6	8	6	10	9	4	7	6	6	8	6	9	7	6
SIGNO	+	-	+	+	+	+	+	+	+	0	+	+	-	0	-

N1 (+) = **10** ; N2 (-) = 3 ; N = 10 + 3 = 13 ; por tanto x = 10 ,
y Z= (2*10 – 13)/ √13 = 1'941 que es < t(12, 0'05)=2'179 y por tanto no se puede rechazar H_0. No se han encontrado diferencias.
Las pruebas no paramétricas son menos potentes que las paramétricas. El test de los signos no ha podido encontrar las diferencias que evidenció la prueba anterior.

Ejercicio 17-5
Un sociólogo quiere investigar si una determinada película sobre la delincuencia juvenil puede cambiar la opinión de las personas adultas de la población X. Para ello estudia una muestra de 100 adultos que han visto la película. Les pregunta si ha cambiado su opinión sobre estos chicos. 15 dicen que siguen opinando lo mismo, 59 los ven con más benevolencia que antes y 26 dicen que los ven peor que antes y que hay que castigarlos con más severidad. ¿Cual es la conclusión?

--- Aquí se puede aplicar el test de los signos, ya que tenemos una opinión después de ver la película, que se contrasta con la que tenían antes de verla.Nos dan los signos ya calculados.
Y tenemos como resultados: 59 + , 26 - , 15 0
Por tanto N_1= 59 , N_2= 26 y N=85 (los 15 que piensan igual no cuentan). X vale pues 59
Ho = no hay cambios de opinión

Z = (2 * 59 - 85) / √85 = 3'58 > $c_{0'001}$ = 3'30 y por tanto se rechaza H_0 a ese nivel de significación. La opinión sobre este asunto ha cambiado significativamente, sobre todo en una mayor tolerancia, pero también, aunque menos, en sentido contrario.

3-b : problema de relación
se resuelven como si ambas variables fueran CT por la fórmula nº 14, que veremos más adelante

4) La variable cualitativa tiene más de dos modalidades y los datos son apareados.

Es un problema de contraste de k medias, que se resuelve por ANOVA-2 (cuyo equivalente no paramátrico es el test de Friedman).
ANOVA-2 permite valorar a la vez dos factores. Dos factores sin repeticiones, ya que hay otros modelos de ANOVA en los que para cada combinación de ambos factores hay más de un dato, "repeticiones", y que no veremos en esta asignatura.
En muchas ocasiones sólo uno de los factores es interesante. El otro, que suelen ser los individuos, pocas veces es objeto de estudio, ya que se sabe de sobras que los individuos difieren bastante entre ellos. Aunque a veces sí puede ser de interés. En todo caso el análisis conjunto es esencial, pues tiene en cuenta la interacción entre ambos factores. Si no se tiene en cuenta esta interacción, el análisis del factor "principal" puede resultar falseado.
Como siempre H_0 dice que no hay diferencias entre las *k* muestras comparadas ni entre los *n* niveles del otro factor, que suelen ser los individuos.
La decisión estadística se toma tras contrastar Z con una F de referencia.
El ANOVA-2 se puede calcular con más facilidad utilizando la siguiente plantilla

Factor A (muestras)

A→	1		2		3		……		k			
Factor B Individuos↓ o bloques	X	X^2	X	X^2	X	X^2	X	X^2	X	X^2	ΣX_A	$(\Sigma X_A)^2$
1												
2												
3												
4												
…..												
n												
											↑ $\Sigma\Sigma X = B$ ↓	↑ **$\Sigma(\Sigma X_A)^2 = C$**
ΣX_B												
n												← $\Sigma n = kn = N$
$\bar{X}$												
$(\Sigma X_B)^2$												← **$\Sigma(\Sigma X_B)^2 = A$**
ΣX^2												← **$\Sigma \Sigma X^2 = D$**

$$C_A = \frac{A}{n} - \frac{B^2}{N} = \square \qquad C_T = D - \square$$

$$C_B = \frac{C}{k} - \frac{B^2}{N} = \square \qquad C_R = C_T - (C_A + C_B) = \square$$

$$V_A = \frac{C_A}{k-1} = \square \qquad V_B = \frac{C_B}{n-1} \square \qquad V_R = \frac{C_R}{(k-1)(n-1)} = \square$$

$$Z_A = \frac{V_A}{V_R} = \square \qquad Z_B = \frac{V_B}{V_R} = \square$$

Valoración de A : por F(k-1 ; (k-1)(n-1)). **Valoración de B** : por F(n-1 ; (k-1)(n-1))

El Anova-2 es una prueba muy robusta, por lo que no comprobamos condiciones de aplicación. De oficio no se nos planteará utilizar la prueba no paramétrica correspondiente, que es el test de Friedman.

En el test de Friedman también es conveniente utilizar una plantilla para hacer los cálculos. Esta plantilla tal cual está diseñada sirve para valorar el factor A, muestras. Si se quiere valorar el otro factor, que llamaremos B, intercambiaremos A y B. Es decir, A lo que antes llamábamos "A" le ponemos el nombre de "B" y viceversa. Los datos se introducen ahora en un orden distinto. Y así podremos estudiar lo que inicialmente era "B".
La prueba de Friedman se valora por Chi-cuadrado, con grado de libertad k-1 (muestras-1).

Si se rechaza H_0 hay que aplicar la prueba de los signos. Se ordenar las muestras, de mayor a menor o viceversa
Y se comparan por parejas, empezando por las más dispares, las de los extremos, de forma similar a como veíamos en el Kruskal—Wallis.

A continuación viene un ejercicio que se resolverá tanto por el ANOVA-2 como por la prueba no paramétrica de Friedman.

Ejercicio 17-6

Queremos probar dos productos estimulantes de la memoria, M1 y M2. Diez personas toman en un orden establecido por el azar M1 , M2 y P (placebo) y cada vez se hace un test de memoria. Se obtienen las siguientes puntuaciones:

M1	M2	P
30	31	26
29	21	19
36	35	37
33	32	27
34	31	26
32	29	30
31	38	35
39	21	14
32	23	19
29	26	29

¿Que producto es el mejor?

**Es un problema de contraste de una Vble. CL, PRODUCTO, con 3 modalidades, M1, M2 y P, y otra CT, que es la PUNTUACION en el test de memoria, que se ha obtenido en cada una de estas tres modalidades. Los datos son apareados. La prueba correspondiente es ANOVA-2. Pero a efectos didácticos se resolverá también por el test de Friedman.

H_0 : no hay diferencias entre las 3 muestras comparadas ni entre los 10 niveles del otro factor, los individuos. En este problema el factor interesante son los productos.

Resolución por ANOVA-2 : Utilizaremos la plantilla de que disponemos.

ANOVA-2 Factor A (muestras)

A→	1		2		3			
Factor B Individuos↓ o bloques	X	X^2	X	X^2	X	X^2	ΣX_A	$(\Sigma X_A)^2$
1	30	900	31	961	26	676	87	7569
2	29	841	21	441	19	361	69	4761
3	36	1296	35	1225	37	1369	108	11664
4	33	1089	32	1024	27	729	92	8464
5	34	1186	31	961	26	676	91	8281
6	32	1024	29	841	30	900	91	8281
7	31	961	38	1444	35	1225	104	10816
8	39	1521	21	441	14	196	74	5476
9	32	1024	23	529	19	361	74	5476
10	29	841	26	676	29	841	84	7056
							874	77844
							↑ $\Sigma\Sigma X$= B ↓	↑ **$\Sigma(\Sigma X_A)^2$= C**
ΣX_B	325		287		262		874	
n	10		10		10		30	← Σn=kn= N
$\bar{X}$	32’5		28’7		26’2			
$(\Sigma X_B)^2$	105625		82369		68644		256638	← **$\Sigma(\Sigma X_B)^2$= A**
ΣX^2		10653		8543		7334	26530	← **$\Sigma\Sigma X^2$= D**

$C_A = \frac{A}{n} - \frac{B^2}{N} =$ 201’2667 $\quad C_T = D - \frac{B^2}{N} =$ 1067’4667

$C_B = \frac{C}{k} - \frac{B^2}{N} =$ 485’4667 $\quad C_R = C_T - (C_A + C_B) =$ 380’7333

$V_A = \frac{C_A}{k-1} =$ 100’6335 $\quad V_B = \frac{C_B}{n-1}$ 53’9407 $\quad V_R = \frac{C_R}{(k-)(n-1)} =$ 21’1519

$Z_A = \frac{V_A}{V_R} =$ **4’76** $\quad Z_B = \frac{V_B}{V_R} =$ **2’55**

Valoración de A : por F(k-1 ; (k-1)(n-1)). **Valoración de B** : por F(n-1 ; (k-1)(n-1))

Sólo nos interesa valorar el factor A, los 3 productos : $Z_A > F(2 ; 18; 0’05)=3’65$ y por tanto se rechaza H_0 a ese nivel de significación: en su conjunto las 3 muestras se comportan de manera significativamente distinta.
Esto nos obliga a realizar la prueba de Scheffé, fórmula 8 bis

El orden es	muestras	1	2	3
	Medias	32’5	28’7	26’2

Comparando 1 y 3 : $Z_{SCH} = (32’5-26’2)^2 / 21’1519*2*0’2=4’69 > F_{0’05}$ .Rechazo de H_0 : 1 es mejor que 3
Comparando 1 y 2 : Z_{SCH} =1’70 <$F_{0’05}$ y no hay rechazo de H_0
Comparando 2 y 3 : Z_{SCH} =0’74 <$F_{0’05}$ y no hay rechazo de H_0
Conclusión final : La prueba de ANOVA-2 nos dice que las puntuaciones de memoria son significativamente distintas en su conjunto. La prueba de Scheffé nos aclara que ello se debe fundamentalmente a la superioridad del producto 1 sobre el 3.

Ahora, el mismo ejercicio resuelto por la prueba no paramétrica, utilizando su plantilla

TEST DE FRIEDMAN

Valoración del factor A

los datos se ordenan por filas

Factor A (muestras)

A→	1		2		3	
B↓ Individuos o bloques	X	R	X	R	X	R
1	30	2	31	3	26	1
2	29	3	21	2	19	1
3	36	2	35	1	37	3
4	33	3	32	2	27	1
5	34	3	31	2	26	1
6	32	3	29	1	30	2
7	31	1	38	3	35	2
8	39	3	21	2	14	1
9	32	3	23	2	19	1
10	29	2'5	26	1	29	2'5
Σ R	25'5		19		15'5	
(Σ R)²	650'25		361		240'25	

1251'5

Σ (Σ R)²

Fórmula: $$Z = \frac{12\sum(\sum R)^2}{nk(k+1)} - 3n(k+1)$$

Valoración de A : por χ2 con g.l. k-1

Z=5'15 < χ2 (2 ; 0'05)=5'99 y por tanto no hay rechazo de H_0 . La prueba no paramétrica, menos potente, no ha podido descubrir las diferencias que sí encontró el ANOVA-2

Valoración de B : (aquí no interesa) ; si interesara, se intercambian los nombres de A y B, es decir, que lo que antes era A pasa a ser B y viceversa y se ponen los datos en la tabla

TEMA 18 : CONTRASTE DE DOS VARIABLES CUANTITATIVAS

Para estudiar la relación o dependencia entre dos variables cuantitativas se valora estadísticamente ("se contrasta") el coeficiente de correlación. En principio los datos son apareados.

Hay una prueba paramétrica , que contrasta el coeficiente de correlación de Pearson, **r** , y otra no paramétrica, que contrasta el coeficiente de correlación de Spearman , $\mathbf{r_s}$

1) PRUEBA PARAMETRICA : Contraste de r

$$\mathbf{Z = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}} \qquad \text{(Fórmula nº 14)}$$

Condiciones de aplicación: Si la muestra es pequeña, igualdad de varianzas de x e y

$$V = \frac{S^2 \text{ mayor}}{S^2 \text{ menor}} < F(N-1, N-1 , 0.05)$$

Valoración : a) si N>30, por c de la D.N.
b) si N<30, por t(N-2, α)

Ejemplo: Ejercicio 18-1
Medimos en 5 sujetos la concentración de cafeína en sangre después de tomar cierta cantidad de café. Al mismo tiempo medimos el tiempo de reacción ante el estímulo H.
Obtenemos:

Individuo	1	2	3	4	5
Cafeína	2	4	3	6	2
Tiempo	11	9	10	7	12

Queremos contestar a la pregunta de si hay o no una relación entre la cafeína en sangre y la rapidez de reflejos.

- Se trata de un problema de contraste entre dos variables cuantitativas : cafeína en sangre y rapidez de reflejos (medida como tiempo de reacción). A resolver por la fórmula nº 14, si cumple la condición de aplicación. Ho: no hay ninguna relación entre las variables, son independientes.
- Hay que comprobar si cumple la condición de aplicación. Para ello tenemos que calcular las respectivas varianzas:

La de la cafeína es 2.8 y la del tiempo de reacción es 3.7
V = 3.7 / 2.8 = 1.32, que es menor que F(4 , 4, 0.05)=6.39 , por lo que sí cumple la condición de aplicación y podemos utilizar la fórmula nª 14, contraste de r

- Calculamos el coeficiente de correlación, y obtenemos **r = -0.979**

  $$\mathbf{Z = \frac{-0,979\sqrt{(5-2)}}{\sqrt{1-(-0,979)^2}}}$$ = -8,32 , que es > t (3, 0.01)=5.84 , por tanto se rechaza Ho a ese nivel de significación. p < 0,01

Existe una relación inversa (signo negativo!) entre cafeína en sangre y rapidez de reflejos: a más cafeína, menor tiempo de reacción (es decir, más rapidez de reflejos), y a menos cafeina, más tiempo de reacción (es decir, reflejos más lentos).

El problema es experimental y por tanto puede establecerse una relación causa-efecto.

2) PRUEBA NO PARAMÉTRICA : contraste de r_s

Es el **test de correlación de rango de Spearman**. Se usa cuando no puede hacerse un contraste de r por no cumplir la condición de aplicación (igualdad de varianzas en el caso de muestras pequeñas) o los datos no proceden de una población distribuída normalmente.
Hay que calcular el coeficiente de correlación de Spearman, r_s , utilizando la siguiente plantilla, que también se ofrece en el cuadernillo de fórmulas.

<u>Fórmula nº 15</u>: Test de rango de Spearman (r_s)

Los datos de X e Y se ordenan por separado

Individ.		**X**	**Y**	**R de X**	**R de Y**		**d**	**d^2**
1								
2								
3								
4								
5								
6								
.....								
N								
							Suma	
								Σd^2

Una vez ordenados los datos se asigna a cada uno de ellos su número de orden (Rango) y se anota en la columna R que corresponda, según se indica más abajo

Cálculo:

$$r_s = 1 - \frac{6\sum d^2}{N\left(N^2-1\right)} \qquad \text{(fórmula nº 15)}$$

Z se calcula por la fórmula 14 , sin condición de aplicación, dándole a r el valor de r_s
Valoración por t_{N-2} (si N<30) ó DN (si N≥30)

<u>Pasos</u>:
1º- ordenar por separado los datos de ambas variables (de mayor a menor o de menor a mayor) , asignándoles números de orden correlativos. Cuando un dato se repite una o más veces (casos "ex equo") a cada uno se le asigna la media de los números de orden que les corresponderían (con un decimal).
2º- se restan los números de orden de cada individuo (d) y esta diferencia se eleva al cuadrado (d^2). Al final, se suma la columna de d^2, obteniendo Σd^2
3º- se aplica la fórmula para calcular **r_s**
4º- r_s se valora por la fórmula nº 14 (poniendo r_s donde dice r) y valorando por la t de Student con g.l. N-2 o por la DN en función de lo que valga N.

<u>Ejemplo</u> . Ejercicio nº 18-2
En 10 individuos realizamos alternativamente al azar un test de memoria y otro de atención, obteniendo las siguientes puntuaciones:

Individuo	1	2	3	4	5	6	7	8	9	10
Memoria	6	4	3	5	3	2	1	5	4	1
Atención	12	6	4	12	6	2	2	14	10	1

¿Están relacionados los resultados de ambos tests?

- se trata de un problema de contraste entre dos variables CT (problema de relación). Ho: no existe ninguna relación. Hay independencia

- A resolver de entrada por la fórmula nº 14, si cumple la condición de aplicación, ya que se trata de una muestra pequeña.
- Calculamos las varianzas, obteniendo 2.93 para M. y 22.77 para A.
  V = 22'7 / 2'93 = 7'77 , que es mayor que F(9, 9 , 0'05)=3'18 ; por tanto no cumple la condición de aplicación y no podemos aplicar la fórmula nº 14, pasando de oficio al test no paramétrico, de la r_s . Para ello, utilizaremos la plantilla, escribiendo primero los datos originales y después calculando los rangos de X (= Memoria) e Y (=Atención)

Individuos	X	Y	R de X	R de Y	d	d^2
1	6	12	10	8.5	1.5	2.25
2	4	6	6.5	5.5	1	1
3	3	4	4.5	4	0.5	0.25
4	5	12	8.5	8.5	0	0
5	3	6	4.5	5.5	-1	1
6	2	2	3	2.5	0.5	0.25
7	1	2	1.5	2.5	-1	1
8	5	14	8.5	10	-1,5	2.25
9	4	10	6.5	7	-0.5	0.25
10	1	1	1.5	1	-0.5	0.25
					$\sum d^2 =$	8.5

- Calculamos $\mathbf{r_s}$:

$$r_s = 1 - \frac{6*8,5}{10\left(10^2-1\right)} = 0,948$$

- Aplicamos la fórmula 14 con r = 0'948 y obtenemos

$$\mathbf{Z = 0,948\sqrt{10-2} / \sqrt{1-0,948^2} = 8,42}$$

que es mayor que t(8, 0'001)=5'041 rechazando a ese nivel de significación la hipótesis nula y aceptando la alternativa con p<0,001. Por tanto , sí hay relación entre ambas puntuaciones. La relación es directa: a mayor nota en Memoria, mayor en Atención y viceversa.

Recordatorio

Con datos independientes los problemas de relación se resuelven con las mismas fórmulas que los de comparación; en cambio si los datos son apareados, las fórmulas son distintas para cada caso.

Según el enfoque que se haga del problema, pueden surgir dudas en algunos casos entre si hay que tomar la fórmula nº 10 ó la nº 14. Ambas sirven para datos apareados. La 10 para el contraste de una variable CL y otra CT, **comparando** los datos CT obtenidos en cada modalidad de la variable CL. En la nº 14 se estudia la **relación**. Por tanto si surge la duda al tratar datos apareados de este tipo sobre la fórmula a usar, preguntarse siempre si es un problema de comparación o de relación. ¿Se pide si los valores son más o menos iguales o no, o bien se pide que se pruebe si hay o no una relación entre ellos?

Si con los datos del ejercicio 18-2 se nos preguntara: ¿Hay diferencias importantes entre los resultados de ambos tests? se trataría del contraste de una variable CL (tipo de test: M y A) y de otra CT (puntuación obtenida en el test, que medimos en ambas modalidades de la CL). Datos apareados. A resolver por la fórmula nº 10, que no tiene condición expresa de aplicación. Ho: no hay diferencias entre las puntuaciones obtenidas en cada test.

Aplicando la fórmula 10 obtenemos : z= -3.416 que es mayor en valor absoluto que t(9, 0.01) que vale 3.250. Por tanto, se rechaza H0 a ese nivel de 0.01 . Sí **hay diferencias significativas** entre las puntuaciones de Memoria y Atención, en el sentido de que las de Memoria son más bajas con p< 0,01

Tema 19 : Demografía sanitaria

MEDIDAS DE LA ENFERMEDAD

En Epidemiología se estudia con detalle la frecuencia de enfermedades, su evolución a curación, cronicidad o muerte y su asociación con determinadas circunstancias : factores de riesgo, factores de pronóstico, tratamientos (medida de la eficacia; efectos secundarios...), forma de vida, medio ambiente, prevención, etc. etc., siguiendo estrategias que verán Uds. en esa asignatura. La estadística es una herramienta básica en Epidemiología, que es sin duda su aplicación más importante en las Ciencias de la Salud.

Aquí vamos a ver algunos índices básicos y su forma de calcularlos.

Prevalencia

Es la proporción de individuos que en un momento dado o en un periodo de tiempo determinado presentan el suceso que se está estudiando.
P = nº sucesos / nº total individuos .
Puede expresarse también como porcentaje o como tasa.

> Si en la ciudad X, que tiene 50 000 habitantes, hay en el año A 1000 personas diabéticas, la prevalencia será :
> P = 1000 / 50000 = 0,02   (ó el 2% , si se prefiere)

También suelen calcularse los intervalos de confianza
.

Incidencia

Es la proporción de nuevos casos (aparición del suceso en nuevos individuos) en un periodo de tiempo determinado, generalmente un año.
I = nº sucesos nuevos / nº total individuos
Si en esa ciudad X en el año A 100 personas se hicieron diabéticas, I = 100 / 50000 = 0,002
Que también se puede expresar como 0,2 % ó 2 ‰ , etc. O como 20 por 10.000 habitantes, o 200 por 100.000 habitantes..
También suelen calcularse los intervalos de confianza.
Hay otras formas de medir la incidencia en las que no entramos aquí.

Las **Odds Ratios,** el **Riesgo Relativo , (**ya vistos), y el **NNT,** que veremos, sirven también para “medir” enfermedades y otros sucesos sanitarios.
Los estudios **caso-control** son herramientas habituales y también los estudios de **cohortes** (de realización más difícil).

Tasas Sanitarias .

Las TASAS son frecuencias relativas referidas a un número preestablecido de individuos, múltiplo de 100. Esto se hace para evitar tasas menores de 1, a veces con varios ceros antes del primer dígito significativo, lo que las haría de difícil manejo. Es mejor una tasa expresada como 5,4 por mil, que como 0,0054

Las TASAS SANITARIAS hacen referencia a fenómenos relacionados con la Sanidad en una población. Hay multitud de ellas. La mayoría reflejan las incidencias naturales

de la población, como las tasas de natalidad, morbilidad, mortalidad, crecimiento vegetativo, etc. Suelen ir referidas al año natural. Como la población varía continuamente a lo largo del año, suele tomarse la que hay (o se estima) el 1 de julio. Como ejemplo se dan algunas de ellas:

TASA DE NATALIDAD: nacimientos en el año dividido por la población y multiplicado por mil: 1000N/P ‰

TASA DE MORTALIDAD GENERAL: defunciones en el año dividido por la población y multiplicado por mil: 1000D/P ‰

Además hay tasas de mortalidad por enfermedades o grupos de enfermedades, sexo, grupos de edad, etc.

TASA DE MORTALIDAD INFANTIL: defunciones de niños menores de un año dividido por el nº de nacimientos vivos en ese año y multiplicado por mil.

$TMI = 1000D_{<1año} / Nv$ ‰

TASA DE CRECIMIENTO VEGETATIVO: nacimientos menos defunciones, dividido por la población y multiplicado por mil. TCV = 1000(N-D)/P ‰

TASA DE ENVEJECIMIENTO: población mayor de 65 años dividido por la población menor de 15 años y multiplicado por cien:

$TE = 100 * P_{>65a} / P_{<15a}$ %

Ejemplos

con datos de la Comunidad Valenciana en el año 2.000

Datos básicos: Población: 4.039.115, de ellos 604.987 menores de 15 años y 682.837 mayores de 65 años.

Hubo 42.046 nacimientos (vivos) y 37.979 defunciones (143 menores de 1 año).

TASA DE NATALIDAD

1000*42.046 / 4.039.115 = 10,4 ‰

TASA DE MORTALIDAD GENERAL

1000*37.979 / 4.039.115 = 9,4 ‰

TASA DE MORTALIDAD INFANTIL

1000*143 / 42.046 = 3,4 ‰

TASA DE CRECIMIENTO VEGETATIVO

1000*(42.046-37.979) / 4.039.115 = 1,007 ‰

TASA DE ENVEJECIMIENTO

100*682.837 / 604.987 = 112,9 %

Indices Hospitalarios

Valoran de forma cuantitativa el trabajo realizado en un Hospital. La valoración de la calidad mediante índices estadísticos está poco desarrollada, dadas sus dificultades.

Se registran los ingresos, las altas y las estancias de todo el hospital y de cada uno de sus Servicios y Unidades. Además se calculan índices que relacionan estos datos con el número de camas. los cómputos pueden hacerse para un solo día, un mes o todo el año.

Un pequeño glosario de los términos más habituales:
Se considera como CAMA HOSIPTALARIA aquella que está montada para su uso regular las 24 horas del día. No se contabilizan como tal las posibles camas del Servicio de Admisión, las de Paritorios, las de Recuperación y otras similares, que ocupan de forma transitoria pacientes que ya tienen su cama en otro lugar.
La suma total de camas hospitalarias de un Hospital o de un Servicio da su CAPACIDAD ACTUAL o REAL.
Se considera que hay INGRESO cuando se ha abierto la correspondiente ficha y el paciente es internado.
Se contabiliza una ESTANCIA cuando el paciente pernocta (está a la “hora censal” , la medianoche) o ha efectuado una de las dos comidas principales.
Cuando se cierra la Historia Clínica y la ficha de ingreso y el paciente abandona su cama (vivo o muerto) se produce el ALTA.
El INDICE O PROMEDIO DE OCUPACION resulta de dividir el nº de estancias multiplicado por cien entre el nº de días y el nº de camas. Es un %
La ESTANCIA MEDIA O PROMEDIO DE ESTANCIA se calcula dividiendo el nº de estancias por el nº de altas.
El INDICE DE ROTACION ENFERMO-CAMA, nº de pacientes que han pasado por una cama en el periodo de tiempo considerado, es igual al cociente del nº de ingresos y el nº de camas.
El INDICE o INTERVALO DE REOCUPACION, tiempo medio que pasa (en días) desde que una cama queda libre hasta que es ocupada de nuevo, es igual al nº de camas por el de días, menos el nº de estancias, todo ello dividido por el nº de altas
Se pueden calcular también promedio de ingresos, de altas, nº de operaciones, de análisis, de radiografías, endoscopias, resonancias magnéticas, etc. etc.

El estudio detallado de estos y otros muchos índices y datos corresponde a otras asignaturas. Aquí se da un esbozo previo para ver la mecánica de los cálculos. Como ejemplo, se van a ver algunos de estos datos e índices para la actividad de hospitalización del Hospital X el año pasado.
Capacidad real del Hospital : 545
Ingresos : 15.768
Altas : 15.752
Estancias : 137.078
Indice o promedio de ocupación : = 100*137.078 / 365 / 545 = 68,9 %
Estancia media o promedio de estancia : 137.078 / 15.752 = 8,7
Indice de rotación enfermo-cama : 15.768 / 545 = 28,93
Indice o Intervalo de reocupación : (545*365 - 137.078) /15752 = 3,98

Análisis de supervivencia

El tiempo que transcurre desde la aparición de un evento hasta la muerte de una persona puede ser de interés en situaciones muy diversas. Por ejemplo:
----- ¿Cuántos años de vida es de esperar que alcance un recién nacido sano?
----- ¿Cuantos años de vida le quedan en media a una persona de X años de edad?
-muy importante para las compañías de seguros (y para el interesado!)-
----- ¿Supervivencia de los pacientes de cáncer?
----- ¿Supervivencia de trasplantados (corazón, riñón, higado...)?

La respuesta a estas cuestiones es el llamado análisis de supervivencia, que se refleja en las **tablas de vida**, también llamadas **tablas de mortalidad** y **tablas de supervivencia**. Es un análisis muy complicado, que iniciaron en el siglo XVII Graunt y Halley (el del cometa) y que se ha ido perfeccionando con el tiempo, convirtiéndose en una especialidad de la Bioestadística. Que depende hoy día totalmente de la informática. Aquí sólo podemos hacer un pequeño esbozo del mismo.

Por extensión, se utiliza este método para situaciones en las que no existe un riesgo de muerte. Por ejemplo, para valorar la eficacia de varios tratamientos del mismo proceso (generalmente enfermedades crónicas) la muerte se sutituye por la recaída y se contabilizan las probabilidades de recaer o seguir asintomático con cada uno de ellos.

Cuando el tiempo es corto, hasta 5 ó 10 años, se habla de **tablas actuales** y cuando es muy prolongado, de **tablas de cohortes**. En ambos casos el tiempo total T se divide en intervalos o periodos iguales, que en función del caso concreto pueden ser días, semanas, meses o años. En cada uno de ellos se anotan los individuos vivos al principio del intervalo, los que mueren en el mismo y los que se pierden del seguimiento (por no estar localizables o haber muerto por otra causa). Y se calculan, entre otras cosas, las probabilidades de morir y sobrevivir en el intervalo. Se puede estudiar a la población en general o a grupos específicos, como hombres, mujeres, diabéticos, fumadores, cancerosos, trasplantados, operados de by-pass, etc, etc

Los medios de comunicación informan a menudo de la **esperanza de vida al nacer**: "Los nacidos el año pasado en España tienen una esperanza de vida de 85 años en mujeres y 78 en hombres" . Son los años que es de esperar que vivan por término medio. El pronóstico sigue la campana de Gauss de la DN; los valores alrededor de la media son los más frecuentes, pero también hay valores extremos, por arriba y por abajo, que, aunque sean poco frecuentes, también se dan. En los países desarrollados estas tablas son muy fiables. Los intervalos son anuales y se puede ver la **expectativa de vida para cada edad**. Por ejemplo en España (datos de 2009): una mujer de 50 años puede esperar 35 años más de vida y una de 90 años 5 más. Si un varón cumple 100 años, su esperanza futura es de 2,75 años más. Para edades inferiores a 40 años la esperanza de vida restante es : 85- edad (mujeres) y 78 - edad (hombres).

En enfermos de cáncer y trasplantados se usan mucho las tablas (y gráficos) de supervivencia.
Como no todos los pacientes enferman a la vez, el cómputo es complicado y muy engorroso, incluso con la ayuda de programas informáticos.

Como muestra un pequeño ejemplo, tomado de De Mould, Clinical Radiology, 1976; 27: 33

Se trata del seguimiento de 150 pacientes de un determinado tipo de cáncer .

intervalo (i) años	casos al inicio	muertes en i	perdidos en i	casos útiles para el cálculo	p de morir en i	p de sobrevivir en i	p total de sobrevivir
1º	150	39	4	148	0,263	0,737	0,737
2º	107	19	2	106	0,179	0,821	0,605
3º	86	12	1	85,5	0,140	0,860	0,520
4º	73	6	1	72,5	0,082	0,918	0,477
5º	66	6	0	66	0,090	0,910	0,434
6º	60	5	1	59,5	0,084	0,916	0,397
7º	54	3	2	53	0,056	0,944	0,374
8º	49	1	1	48,5	0,020	0,980	0,366
9º	47	3	4	45	0,066	0,934	0,341
10º	40	2	4	38	0,052	0,948	0,323
	34						

Al final del 10º intervalo quedan en seguimiento 34 pacientes.

Para hallar los casos útiles para el cálculo se ha restado de los casos al inicio del intervalo la mitad de los casos perdidos. Ya que se asume que se han distribuido uniformente a lo largo del periodo y por tanto en media han estado medio intervalo expuestos al riesgo de morir.

Los casos al inicio de cada periodo se obtienen restando a los del periodo anterior los muertos y perdidos.

La probabilidad de morir en el intervalo 1º es 39/148 = 0,263513, mal redondeado a 0,263 ; por tanto la de sobrevivir es 1-0,263 = 0,737. En los restantes intervalos se hacen cálculos similares.

La probabilidad total de supervivencia es para el primer intervalo también 0,737 . Para los demás se obtiene multiplicando la p de sobrevivir en ese intervalo por la total del intervalo anterior; así para el 6º intervalo la p total es 0,916*0,434 = 0,397 (recordar la ley multiplicativa: probabilidad de haber llegado a este intervalo **y** probabilidad de sobrevivir a este intervalo)

NNT

NNT = Number Need to Treat o **número necesario a tratar**. Es el número de individuos que hay que tratar con el tratamiento experimental para evitar un evento desfavorable o para conseguir un efecto favorable. Como referencia hay un grupo control. Por ejemplo, se puede recomendar con la intención de evitar una enfermedad que todas las personas que reúnan ciertas condiciones tomen un determinado medicamento, que vale su dinero y puede dar efectos secundarios. El tiempo y los estudios nos dirán si es eficaz y en caso positivo cuantas personas hay que tratar para evitar un caso de enfermedad o muerte: 10, 200 ó 5000 o lo que sea. Valorando los

efectos secundarios, económicos y de todo tipo que tiene esa recomendación se podrán sacar las consecuencias oportunas.

Se calcula así (utilizando los términos genéricos de una tabla 2x2) :
(colocamos en primer lugar , en a_1, b_1 y N_1, los datos de los controles) :

$$NT = \frac{1}{\frac{a_1}{N_1} - \frac{a_2}{N_2}}$$

Ejemplo: Se da diariamente el medicamento M con la intención de evitar el evento E a 3051 personas y se controla también a 3054 personas que no toman el medicamento. Pasados 5 años 307 de los que tomaron M presentaron el evento E, por 420 de los que no lo tomaron. Calcular el NNT

	Controles	Tratados
Evento E +	420	307
Evento E -	2634	2744
	3054	3051

NNT = 1 / (420/3054 - 307/3051) = 27,1 ≈ 27

O sea que por cada 27 pacientes tratados con el medicamento M se evitaría un evento E
Los expertos tendrán que valorar si lo que se hace es buena estrategia: dependerá de la naturaleza del evento a evitar, del coste del medicamento, de sus efectos secundarios, etc.

Tema 20 : Errores de las medidas de laboratorio. Control de calidad. Valoración de pruebas diagnósticas.

Errores analíticos

Los análisis son muestreos que nos informan a partir de un pequeño espécimen de lo que ocurre en toda la sangre o en toda la orina o en todo el líquido cefalorraquídeo, etc. (que son la población). El resultado de un análisis es el valor puntual de una estimación a partir de la muestra. Por tanto, como toda estimación, los análisis tienen su error muestral inevitable. Sería deseable que los resultados se dieran también con su intervalo de confianza.

Errores en recuentos y porcentajes

**para un recuento : $\mathbf{cf\sqrt{N}}$

**para un porcentaje : $\mathbf{c\sqrt{\frac{pq}{N}}}$

siendo c la nota tipificada de la DN ; para α = 0'05 vale 1'96
f el factor de multiplicación del método
N el nº de elementos realmente contados

Ejemplo en determinaciones sanguíneas.

Tanto en los recuentos clásicos como los que hacen los modernos aparatos sólo se cuenta una parte y luego se multiplica por el llamado factor de multiplicación. (¡sería tarea imposible contar 5.000.000 de hematíes!

Determinación	f	N	Resultado/ml	Error ±
HEMATIES				
en cámara	10.000	500	5.000.000	450.000
		320	3.200.000	350.000
Coulter I	500	10.000	5.000.000	100.000
		6.400	3.200.000	80.000
LEUCOCITOS				
en cámara	100	100	10.000	2.000
		20	2.000	875
Coulter I	2'5	4.000	10.000	310
		800	2.000	140
CELULAS LCR	1/3	36	12	4
		3.600	1.200	40

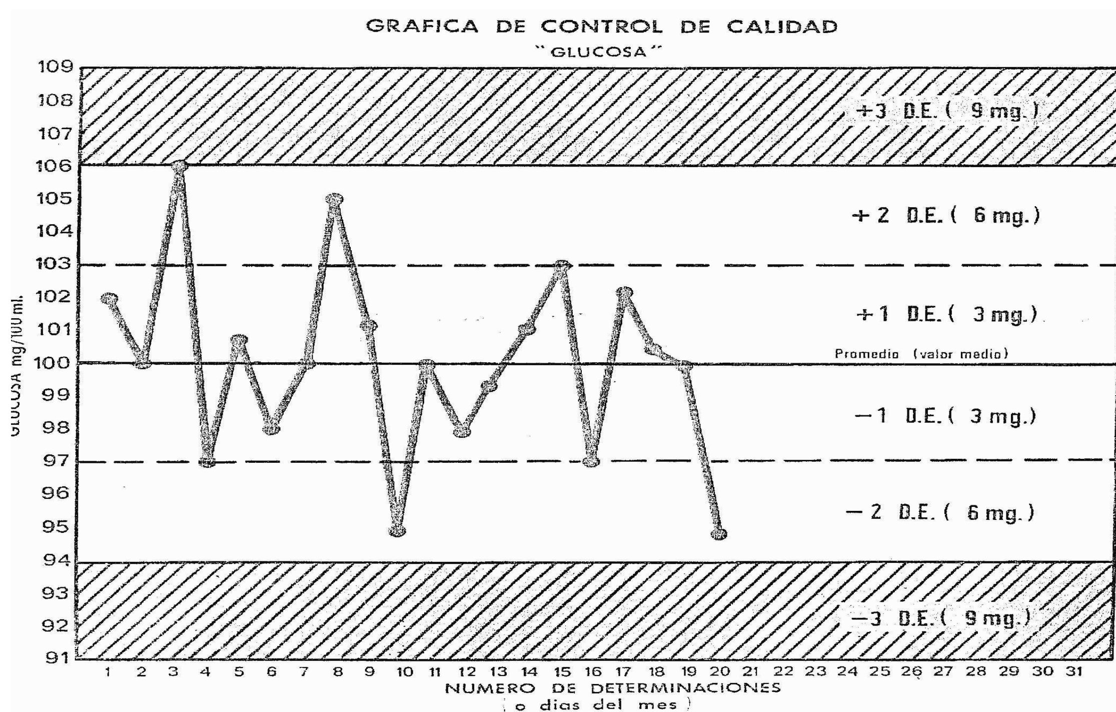
FORMULA LEUCOCITARIA	N	Para un resultado de	Error ±
POLINUCLEARES	200	60 %	7
	100		10
	50		14
EOSINOFILOS	200	6%	3
	100		5
	50		7

Errores analíticos en determinaciones químicas

Se acepta como margen de variación el llamado "**intervalo normal**" : **$x \pm 2s$**
La s (desviación estándar) la fija el fabricante del reactivo en base a sus ensayos.

Control de calidad

Es un sistema para medir la precisión y exactitud de las determinaciones analíticas. mide multitud de factores, como calidad de los reactivos, calidad y puesta a punto de los aparatos, preparación de las muestras, habilidades personales, etc.
Hay varios procedimientos:
--Uno muy sencillo es utilizar un gráfico de control en el que están marcadas desviaciones estándar del método a controlar, con una zona central de variaciones aceptables y otras periféricas, que indican error importante. Se hace cada día una determinación con un patrón de control, de composición conocida, y se pasa al gráfico.



--Otro procedimiento es hacer la determinación por duplicado y comparar los resultados.
--A los modernos aparatos se les debe pasar cada día patrones de control, que de forma automática, informan de la calidad de las determinaciones.
--El mejor método, al que ya se acogen la mayoría de Laboratorios, es el de los controles externos. Centros especiales, de alta tecnología, remiten periódicamente a los Laboratorios asociados muestras para que hagan en ellas las determinaciones que se les piden. Estos devuelven los resultados. El controlador les comunica al cabo de un tiempo los resultados verdaderos , junto a los resultados globales de todos los Laboratorios participantes.

VALORACION DE PRUEBAS DIAGNOSTICAS

Los análisis y pruebas diagnósticas, de cribado o no, pueden ser valorados calculando varios parámetros, que veremos de la mano de un supuesto.
Hacemos una prueba para ver si alguien está o no enfermo.
Si sale positivo (P) puede ser positivo verdadero (PV) o falso positivo (PF).
Si sale negativo (N) puede ser negativo verdadero (NV) o falso negativo (NF).
Un sano (Sa) puede dar positivo o negativo y un enfermo (En) también.

Valoración pruebas		Prueba		
		+	-	
Enfermedad	+	PV	NF	**En**
	-	PF	NV	**Sa**
		P	**N**	Total

o bien

Valoración pruebas		Enfermedad		
		+	-	
Prueba	+	PV	PF	**P**
	-	NF	NV	**N**
		En	**San**	Total

Sensibilidad (S) = PV*100/En
Especificidad (E) = NV*100/Sa
Eficiencia de la prueba (EP) = (PV+NV)*100/Total
Valor predictivo resultado + (VPRP) = PV*100/P
Valor predictivo resultado - (VPRN) = NV*100/N
Cociente de probabilidades + (CP+) = S/(100-E)
Cociente de probabilidades - (CP-) = (100-S)/E

Si se trabaja con porcentajes, aparece el 100 en la fórmula. Usando proporciones, en vez del 100 hay que poner 1

La sensibilidad es la positividad en la enfermedad , el % de positivos entre los enfermos

La especificidad es la negatividad en salud, el % de negativos entre los sanos

Valor predictivo de un resultado positivo es el % de positivos que están realmente enfermos

Valor predictivo de un resultado negativo es el % de negativos realmente sanos

Eficiencia de la prueba : el % de diagnósticos correctos

El cociente de probabilidades de una prueba positiva o cociente de verosimilitud + (CP+), (también muy conocido por su nombre en inglés : likelihood ratio of positive test) es el cociente de las probabilidades de positivos verdaderos y falsos positivos (aunque no lo parezca por su fórmula). Suele expresarse como frecuencia relativa, no como % .

El cociente de probabilidades de una prueba negativa o cociente de verosimilitud (CP-) (su nombre en inglés : likelihood ratio of negative test) es el cociente de las probabilidades de falsos negativos y negativos verdaderos. Suele expresarse como frecuencia relativa, no como % .
Estos cocientes son mejores índices que los valores predictivos , ya que éstos depende de la proporción de enfermos en la muestra (de la prevalencia) y los CP no. Sólo dependen de la sensibilidad y de la especificidad. Permiten comparar métodos diagnósticos diferentes y valorar si la probabilidad pre-prueba cambiará tras conocerse el resultado del análisis. Las CP están muy cerca de 1, cuando apenas varía la p pre-prueba. Al alejarse de 1 aumenta la variación. (Lo veremos enseguida)

La siguiente tabla nos puede orientar sobre la variación que ocurrirá:
Cambios esperados de la probabilidad pre-prueba según el valor de las CP

CP+	1	< 5	5 a 10	> 10
CP-	1	> 0’2	0’1 a 0’2	< 0’1
	No cambia	Cambio escaso	Cambio moderado	Cambio intenso

En todos estos parámetros también se calculan intervalos de confianza (IC), lo que es muy fácil para S y E, ya que son proporciones. (¡Ojo! Para calcular el IC de S hay que tomar N=En y para el de E, N=Sa) . Para el resto de índices el cálculo es más complejo y pasamos de ello.

Ejemplo de cálculos:

		Enfermedad		
		+	-	
Prueba	+	72	100	172
	-	18	150	168
		90	250	340

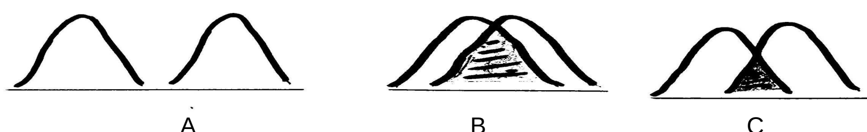
S = 80%, con IC $\in$(71,7%÷88,3%)
E = 60 % con IC $\in$(53,9%÷66,1%)
Eficiencia = 65,3%
VPRP = 41,9%
VPRN = 89,3%
CP+ = 2
CP- = 0'33
Estos resultados también pueden expresarse como frecuencia relativa (0,8 ; 0,6 ....)!!!

Una buena prueba debe tener S y E lo más cerca posible de 100% (ó de 1). Como mínimo 90% (ó 0,90)
La OR vale en este caso 6'0 con un IC que va de 3'38 a 10'66. Como excluye a 1 es significativo: el análisis + eleva significativamente el riesgo de padecer la enfermedad (y viceversa).
(Recordar que la OR va referida siempre a la casilla a_1 (PV) ; si OR>1 : asociación positiva ; si es <1 , negativa)

Probabilidad pre-prueba y post-prueba
Una persona antes de someterse al test tiene una cierta probabilidad de estar enfermo (probabilidad pre-prueba = P_{pre}). Se estima así: P_{pre} = P/N . En el ejemplo: 90/340 =0,265 , que es la prevalencia (mejor expresada como % : 26,5%)
Si sale + , aumenta su probabilidad de estar enfermo y si sale negativo, aumenta su probabilidad de estar sano. Es la llamada probabilidad post-prueba (P_{post}). Se puede calcular a partir de los datos de la tabla y también a partir de P_{pre} y del CP correspondiente.
---a partir de la tabla: Si ha salido + : P_{post} =PV/P ; si ha salido - : P_{post} =NF/N
---a partir de la P_{pre} y de las CP: **$P_{post} = P_{pre} * CP/(1+ P_{pre} (CP-1))$**
Para un resultado + se elige la CP+ y para uno negativo la CP-
En el ejemplo anterior, cuya P_{pre} era de 0,265 :
Si ha salido + : P_{post} = 72/172 = 0,419 ó P_{post} = 0'265*2 / ((1+0'265(2-1))=0'420
(la p de estar enfermo sube del 26% al 42%)
Si ha salido - : P_{post} = 18/168 = 0,107 ó P_{post} = 0'265*0'33 / ((1+0'265(0'33-1)) =0'106
(la p de estar enfermo baja del 26% al 11%)
La probabilidad previa cambia al tener el resultado del análisis.
Según la "predicción" de la tabla de CP (pág.20-3) eran de esperar "cambios escasos",

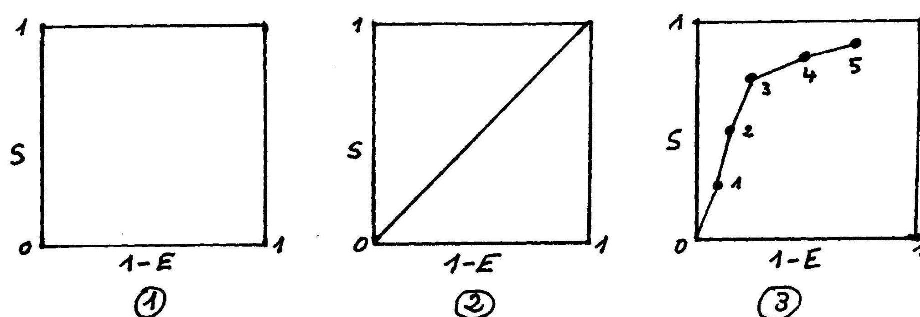
CURVAS ROC

Las pruebas diagnósticas ayudan a diferenciar entre dos (a veces más) situaciones (sano / enfermo, repuesta al tratamiento / no respuesta, etc.). Esto conlleva la posibilidad de error, ya que puede haber falsos positivos y falsos negativos, debido a que casi siempre hay una zona de transición, de solapamiento. El problema está en buscar un punto de corte, un nivel de decisión que permita colocar a cada caso concreto en un sitio o en otro, minimizando la posibilidad de error. En unos caso interesa buscar un límite superior: por ejemplo un valor de glucemia a partir del cual una persona es considerada como diabética. Si se cuentan espermatozoides, se busca un límite inferior que indique esterilidad. El potasio sérico tiene un intervalo, que sobrepasado por arriba o por abajo pone en peligro la vida del individuo y requiere un tratamiento inmediato y adecuado.



En A no hay solapamiento y por tanto discrimina perfectamente. En B hay tal solapamiento, que no discrimina nada. En el caso C, el más frecuente, hay un solapamiento parcial y hay que buscar un buen punto de corte que discrimine con el mínimo error. Esto se puede hacer construyendo una **curva ROC** (siglas del nombre en inglés de Receiver Operating Characteristic, que se puede traducir por Característica con que Opera el Receptor. Esta terminología procede de los primeros tiempos del radar y los controladores dudaban si lo que veían era o no un avión) .
Las curvas ROC sirven pues para elegir un punto o nivel de corte apropiado. Además permiten valorar el rendimiento global de una prueba (calculando el área bajo la curva) y comparar dos curvas, es decir, dos pruebas. Aquí veremos únicamente la elección del punto de corte.

Hay diversos métodos para elegir el punto de corte. El más sencillo es ir probando con diversos puntos y llevar a un gráfico ROC su sensibilidad (S) en el eje vertical y uno menos la especificidad (1-E) en el horizontal. Es conveniente hacer previamente una tabla en la que estén los valores de S y 1-E para cada punto de corte. (Resulta más cómodo trabajar con la sensibilidad y la especificidad expresados como porcentaje. Entonces 1 – E se convierte en 100 – E)



El **nivel de corte** ideal sería el que nos diera un punto en el ángulo superior izquierdo (S = 1 ó 100%, E = 1 ó 100% y por tanto 1-E = 0) como en el caso 1. Cuando hay un solapa-

miento total obtenemos una línea como en el caso 2 (la curva ROC se ha convertido en una recta, la diagonal). Lo habitual es una curva como en el caso 3, en el que vemos que al aumentar la sensibilidad (S), disminuye la especificidad (E) y por tanto aumenta 1-E. Es decir que mejoramos en una cosa y empeoramos en otra.. **El mejor punto de corte, desde el punto de vista estadístico, será aquel que esté más cerca del ángulo superior izquierdo del gráfico**. (En el ejemplo, el punto nº 3). Aunque puede ser modificado en función de la trascendencia que puede tener una mala clasificación, es decir, los falsos positivos y negativos (por ejemplo es muy importante reconocer **todos** los hipotiroidismos congénitos en la prueba que se hace a los recién nacidos, lo que conlleva que de entrada, al bajar el punto de corte, no se escape ningún enfermo, pero haya bastantes casos sospechosos, que angustian a la familia y luego no se confirman)
La tabla también nos orienta hacia **el mejor punto de corte**.

Será aquel en el que la suma de S y 100-E esté más cerca de 100. (si se utiliza proporción en vez de porcentaje, se substituye 100 por 1)

Ejemplo:
El valor de la CPK en 360 pacientes sospechosos de padecer infarto de miocardio (IM) se distribuyó de la siguiente manera entre los que al final tenían y no tenían IM:

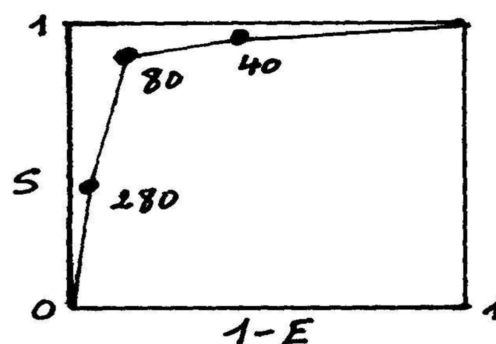
	IM	
CPK ↓	SI	NO
≥280	97	1
80-279	118	15
40-79	13	26
<40	2	88
Total	230	130

En todas las clases hay personas con diagnóstico final de "infarto" y de "no infarto". Para que el análisis sea útil hay que encontrar el punto de corte que mejor clasifique a ambos grupos.
Lo mismo habría que hacer con los otros procedimientos que contribuyen al diagnóstico: síntomas clínicos, electrocardiograma, ecocardiograma, etc.

Se calculan la S y E de los puntos de corte 280 , 80 y 40 y se pasan al gráfico:

Punto de corte de CPK

	280	80	40
S (en %)	**42**	**94**	**99**
E (en %)	99	88	68
100-E	**1**	**12**	**32**
S + (100-E)	*43*	***106***	*131*



El mejor punto de corte parece ser por la tabla y por el gráfico **80**

La tabla siguiente, con los pacientes divididos según el corte, además de permitir calcular S, E , PVP , PVN , CP+ y CP- nos permite hallar la **tasa de probabilidad (TP) ,** en este caso, de ser bien clasificado como IM+ ó mal como IM-

	IM+	IM-	
≥80	215	16	231
<80	15	114	129
Total	230	130	360

Para IM+ sería TP = PV*N / P*NF = 8
Para IM- sería: TP = PF*N / P*NV = 0'078

para estos símbolos ver la tabla de la página 20-3

Cuando no hay discriminación, ambos están alrededor de 1.
Aquí se alejan bastante de 1 y por tanto hay discriminación : el punto de corte elegido parece ser bueno.

Hospital General de Castellón
Servicio de Pediatría

BIOESTADISTICA

Tema 21

PROGRAMAS ESTADISTICOS del CDC de Atlanta

Analysis – Statcalc – Epitable del Epi Info 6

Otros programas:

OpenStat

PSPP

EPI INFO del CDC de Atlanta

El CDC, Centro para el control de enfermedades de Atlanta, tiene un programa basado en DOS, el Epi Info 6 (versión 6.04), Atlanta, cuya difusión es libre y gratuita. Este programa también funciona en Windows. Se puede cargar, en español, en
http://ccp.ucr.ac.cr/cursoweb/epi6.htm

En mi opinión es superior a la versión para Windows, el "Epi Info for Windows", que pasados varios años aún es de manejo difícil, más incompleto y todavía con algunos problemas., aunque también con notables mejoras. La última versión, de agosto de 2008, se puede descargar, en inglés, en la Web del CDC, cuya dirección es : http://www.cdc.gov/epiinfo . Hay también versiones más antiguas en español.

Epi6 tiene otras muchas posibilidades que pueden verse en la AYUDA (F1) o en el detallado MANUAL. En las partes en que no funcione el ratón, utilizar las teclas de dirección (flechas).

Epi 6 tiene varios programas. Sólo nos interesan ANALYSIS , EPITABLE Y STATACALC. Estos programas se pueden descargar también desde http://www.eduardobuesa.es/ , en el subdirectorio Programas

---

ANALYSIS

DATOS:
Trabaja con datos originales, que tiene que estar en un fichero.

a) los ficheros propios tiene la extensión .REC , pero también lee ficheros de dBase III con extensión .DBF
b) EXCEL (de Microsoft Office) permite guardar los ficheros como archivo .DBF, lo que permite generar ficheros legibles para Analysis, si no se dispone del dBase III. El Excel 2007 ya no lo hace, pero sí el Access, al que se pueden pasar los datos desde Excel.
c) El programa sólo guarda los ficheros que han sido cargados con la extensión .REC. Para guardar un fichero cargado como .DBF y que ha sido modificado en el uso del programa hay que reconvertirlo en fichero .REC. Se hace tecleando así:

ROUTE destino:fichero.rec (destino es c: ó d: o la dirección que sea)
WRITE RECFILE
p.e. ROUTE c:\epiestad/biofich.rec , WRITE RECFILE
Si hubiera un fichero con ese nombre hay que borrarlo antes.

Se pueden crear programas, (*.PGM) , con un editor de texto. Hacen automáticamente lo que se ordena. En el programa hay varios ficheros de ejemplo.

Vamos a ver el programa utilizando un fichero que he creado con el nombre BIOEJEMP.REC. Sus datos podrían proceder de 15 personas en las que hemos recogido las siguientes variables: sexo (M, H), categoría laboral o grupo (1 , 2 , 3 , 4), dominio del inglés (S, N), un análisis cuantitativo VALOR1, otro análisis VALOR2, que se repite al cabo de un tiempo VALOR3. Se ha calculado lo que llamamos VALORDIF, que es la diferencia entre VALOR3 y VALOR2.

Los resultados de los cálculos se pueden imprimir, pulsando previamente la tecla F5. Otra opción es abrir un fichero de texto, que se abre con la orden ROUTE y se cierra con CLOSE. (por

ejemplo: ROUTE c:\ficherin.txt). Luego se puede editar con un procesador de textos (Word, Wordpad, etc.). Epi6 tiene uno, muy flojo, llamado EPED.

He recogido los resultados tal como los dan los programas. Como han sido escritos con teclado de USA, no escribe bien las palabras con acentos, ñ y algunos símbolos. He corregido algunos y otros los he dejado tal cual aparecen en pantalla.

El programa utiliza otro lenguaje al que hemos visto en clase. A los resultados de cada prueba los llama como el parámetro de referencia: t de Student, χ^2, F, ... Como es habitual en programas estadísticos no utiliza como referencia la DN, sino exclusivamente la t de Student. Además puede calcular la p de forma continua, no por los hitos de 0,05 , 0, 01 , 0,001 .

El fichero lo creamos con EXCEL según se ve a continuación:

	A	B	C	D	E	F	G	
1	SEXO	GRUPO	INGLES	VALOR1	VALOR2	VALOR3	VALORDIF	
2	H	1	N	12	28	21	7	
3	M	3	N	14	22	20	2	
4	H	2	S	11	21	19	2	
5	H	1	S	18	31	32	-1	
6	H	1	S	16	45	40	5	
7	M	2	N	21	23	16	7	
8	M	4	N	16	28	15	13	
9	H	3	S	27	16	17	-1	
10	M	4	N	29	35	32	3	
11	H	4	S	15	41	32	9	
12	M	2	S	11	39	32	7	
13	M	1	N	21	27	26	1	
14	H	3	S	18	19	12	7	
15	M	2	S	21	20	18	2	
16	H	2	S	15	33	21	12	
17								
18	En FORMATO ajustar la anchura de las columnas a "Autoajustar a							
19	la selección". guardar el archivo en la carpeta en que esté Epi6							
20	como archivo de dBaseIII. Hace varias preguntas: aceptar todo...							
21								

Ya tenemos el fichero como Bioejemp.dbf . Se guarda en la carpeta en que esté Epi6. Lo podemos reconvertir en fichero con extensión REC de la forma que ya hemos visto. Pero si no se van a modificar los datos, no es imprescindible, pues Analysis lo puede leer.

Entramos en ANALYSIS

I.—CARGAR EL FICHERO BIOEJEMP

Teclas importantes:
F1 ayuda , F2 órdenes , F3 variables, etc.

Teclear: READ bioejemp.rec o bioejemp.dbf

READ solo , da un listado de los ficheros REC disponibles. Se puede elegir uno y pulsar.

II –Listado de los datos del fichero

Teclear LIST

REC	SEXO	GRUPO	INGLES	VALOR1	VALOR2	VALOR3	VALORDIF
1	H	1	N	12	28	21	7
2	M	3	N	14	22	20	2
3	H	2	S	11	21	19	2
4	H	1	S	18	31	32	-1
5	H	1	S	16	45	40	5
6	M	2	N	21	23	16	7
7	M	4	N	16	28	15	13
8	H	3	S	27	16	17	-1
9	M	4	N	29	35	32	3
10	H	4	S	15	41	32	9
11	M	2	S	11	39	32	7
12	M	1	N	21	27	26	1
13	H	3	S	18	19	12	7
14	M	2	S	21	20	18	2
15	H	2	S	15	33	21	12

III–Variables cualitativas

A) Frecuencias y porcentajes con intervalo de confianza

Teclear FREQ SEXO /C

SEXO	Frec	Porcent	Acum.	95% Límites Conf
H	8	53.3%	53.3%	26.6%-78.7%
M	7	46.7%	100.0%	21.3%-73.4%
Total	15	100.0%		

B) Contraste de dos variables cualitativas

--con 2 modalidades cada una,datos independientes (tabla de 2x2)

Teclear TABLES SEXO INGLES

SEXO	INGLES N	S	Total
H	1	7	8
M	5	2	7
Total	6	9	15

```
                        Análisis de tabla simple

Odds ratio                                                            0.06
Límites de confianza de Cornfield al 95% de OR          0.00 < OR <   1.21
Estimador de la Máxima Verosimilitud de OR (EMV)                      0.07
Límites de confianza exactos del EMV al 95%             0.00 < OR <   1.16
Límites de Mid-P exactos del EMV al 95%                 0.00 < OR <   0.87
Probabilidad de EMV <=  0.07 si OR poblacional = 1.0            0.03496503

RAZON DE RIESGOS (RR) (Efecto:INGLES=N; Exposici¢n:SEXO=H)            0.17
Límites de confianza al 95% del RR                      0.03 < RR <   1.16

        Ignora la razón de riesgos si es un estudio de casos controles
```

```
                    Chi-Cuadr.   Valores-P
                    -----------  ----------

Sin corregir:            5.40     0.02011616 <---
Mantel-Haenszel:         5.04     0.02474467 <---
Correcci¢n de Yates:     3.23     0.07250203

Test exacto de Fisher: Valor de P para 1 cola: 0.0349650 <---
                       Valor de P para 2 colas: 0.0405594 <---
```

Un valor esperado es < 5; se recomiendan los resultados exactos de Fisher.

-- con más de 2 modalidades en alguna variable (tabla de fxk)
aplica nuestra fórmula nº 3

Teclear TABLES SEXO GRUPO

SEXO \ GRUPO	1	2	3	4	Total
H	3	2	2	1	8
M	1	3	1	2	7
Total	4	5	3	3	15

```
 Un valor esperado es < 5. Chi cuadrado Incorrecto.
      Chi cuadrado =       1.81
Grados de libertad =          3
        Valor de P = 0.61318784
```

IV- Una ó más variables son cuantitativas

a) Estadística descriptiva
Calcula varios parámetros importantes

Teclear FREQ VALOR1 o MEANS VALOR1

VALOR1	Frec	Porcent	Acum
11	2	13.3%	13.3%
12	1	6.7%	20.0%
14	1	6.7%	26.7%
15	2	13.3%	40.0%
16	2	13.3%	53.3%
18	2	13.3%	66.7%
21	3	20.0%	86.7%
27	1	6.7%	93.3%
29	1	6.7%	100.0%
Total	15	100.0%	

Total	Suma	Media	Varianza	Desv est	Error est
15	265	17.667	28.810	5.367	1.386

M¡nimo	Percen.25	Mediana	Percen.75	M ximo	Moda
11.000	14.000	16.000	21.000	29.000	21.000

La T de Student es válida si la media difiere de cero.
Estad¡stico T = 12.748, gl = 14 valor-p = 0.00000
(Esto sirve para aplicar la fórmula nº 10, si ponemos d en vez de VALOR1)

b) Contraste de una variable cualitativa con 2 modalidades y otra cuantitativa; datos independientes.
(= contraste de dos medias = "prueba de la t de Student" = "Unpaired t-test")
aplica nuestras fórmulas nº 6 , 7 , 8 y 9

Teclear MEANS VALOR2 SEXO /N

MEANS de VALOR2 para cada categor¡a de SEXO

SEXO	Observados	Total	Media	Varianza	Desv Est
H	8	234	29.250	107.643	10.375
M	7	194	27.714	49.238	7.017
Diferencia			1.536		

SEXO	M¡nimo	Percen.25	Mediana	Percen.75	M ximo	Moda
H	16.000	20.000	29.500	37.000	45.000	16.000
M	20.000	22.000	27.000	35.000	39.000	20.000

ANOVA
(S¢lo para datos distribuidos normalmente)

Variaci¢n	SC	gl	MC	Estad¡stico F	valor-p	valor-t
Intra	8.805	1	8.805	0.109	0.746408	0.330337
Inter	1048.929	13	80.687			
Total	1057.733	14				

Test de homogeneidad de la varianza de Bartlett's
Chi cuadrado de Bartlett's = 0.878 g. libertad = 1 valor-p = 0.348835

Las varianzas son homog,neas con un 95% de confianza.
Se puede utilizar el ANOVA si las muestras est n distribuidas normalmente.

Test Mann-Whitney o Wilcoxon 2-muestras (test Kruskal-Wallis para dos grupos)

H Kruskal-Wallis (equivalente a Chi cuadrado) = 0.030
Grados de libertad = 1
valor p = 0.862065

El programa ha calculado el ANOVA-1 y el Kruskal-Wallis, aunque sólo hay dos muestras, pero el resultado es correcto. Nuestra Z es aquí "valor-t"

c) Contraste de una variable cualitativa con 3 o más modalidades y otra cuantitativa. Datos independientes (= contraste de 3 ó más medias = ANOVA 1)
calcula ANOVA-1 y Kruskal-Wallis

Teclear MEANS VALOR2 GRUPO /N

MEANS de VALOR2 para cada categor¡a de GRUPO

GRUPO	Observados	Total	Media	Varianza	Desv Est
1	4	131	32.750	69.583	8.342
2	5	136	27.200	70.200	8.379
3	3	57	19.000	9.000	3.000
4	3	104	34.667	42.333	6.506

GRUPO	M¡nimo	Percen.25	Mediana	Percen.75	M ximo	Moda
1	27.000	27.500	29.500	38.000	45.000	27.000
2	20.000	21.000	23.000	33.000	39.000	20.000
3	16.000	16.000	19.000	22.000	22.000	16.000
4	28.000	28.000	35.000	41.000	41.000	28.000

ANOVA

(S¢lo para datos distribuidos normalmente)

Variaci¢n	SC	gl	MC	Estad¡stico F	valor-p
Intra	465.517	3	155.172	2.882	0.084089
Inter	592.217	11	53.838		
Total	1057.733	14			

Test de homogeneidad de la varianza de Bartlett's

Chi cuadrado de Bartlett's = 1.910 g. libertad = 3 valor-p = 0.591212

Las varianzas son homog,neas con un 95% de confianza.

Se puede utilizar el ANOVA si las muestras est n distribuidas normalmente.

An lisis de la Varianza de una v¡a de Kruskal-Wallis

H Kruskal-Wallis (equivalente a Chi cuadrado) = 7.110

Grados de libertad = 3

valor p = 0.068473

d). Contrate de una variable cualitativa con 2 modalidades y otra cuantitativa ; datos apareados. (= contraste de 2 medias con datos apareados = "prueba de de la t de Student para datos apareados" = "paired t-test")

Recordar lo dicho en IV-a: VALORDIF equivale a nuestra d

Teclear FREQ VALORDIF o MEANS VALORDIF

VALORDIF	Frec	Porcent	Acum
-1	2	13.3%	13.3%
1	1	6.7%	20.0%
2	3	20.0%	40.0%
3	1	6.7%	46.7%
5	1	6.7%	53.3%
7	4	26.7%	80.0%
9	1	6.7%	86.7%
12	1	6.7%	93.3%
13	1	6.7%	100.0%
Total	15	100.0%	

Total	Suma	Media	Varianza	Desv est	Error est
15	75	5.000	18.857	4.342	1.121

M¡nimo	Percen.25	Mediana	Percen.75	M ximo	Moda
1.000	2.000	5.000	7.000	13.000	7.000

La T de Student es v lida si la media difiere de cero.

Estad¡stico T = 4.459, gl = 14 valor-p = 0.00054

e) Coeficiente de correlación y ecuación de regresión. contraste de 2 variables cuantitativas.

Teclear REGRESS VALOR 3 VALOR2

Coeficiente de correlaci¢n: r = 0.87 r^2 = 0.76

L¡mit. de confianza al 95%:0.40 < r^2 < 0.91

Fuente	gl	Suma Cuadrados	Media Cuadrados	Estad¡stico-F
Regresi¢n	1	715.1482	715.1482	40.32
Residuales	13	230.5852	17.7373	
Total	14	945.7333		

Coeficientes B

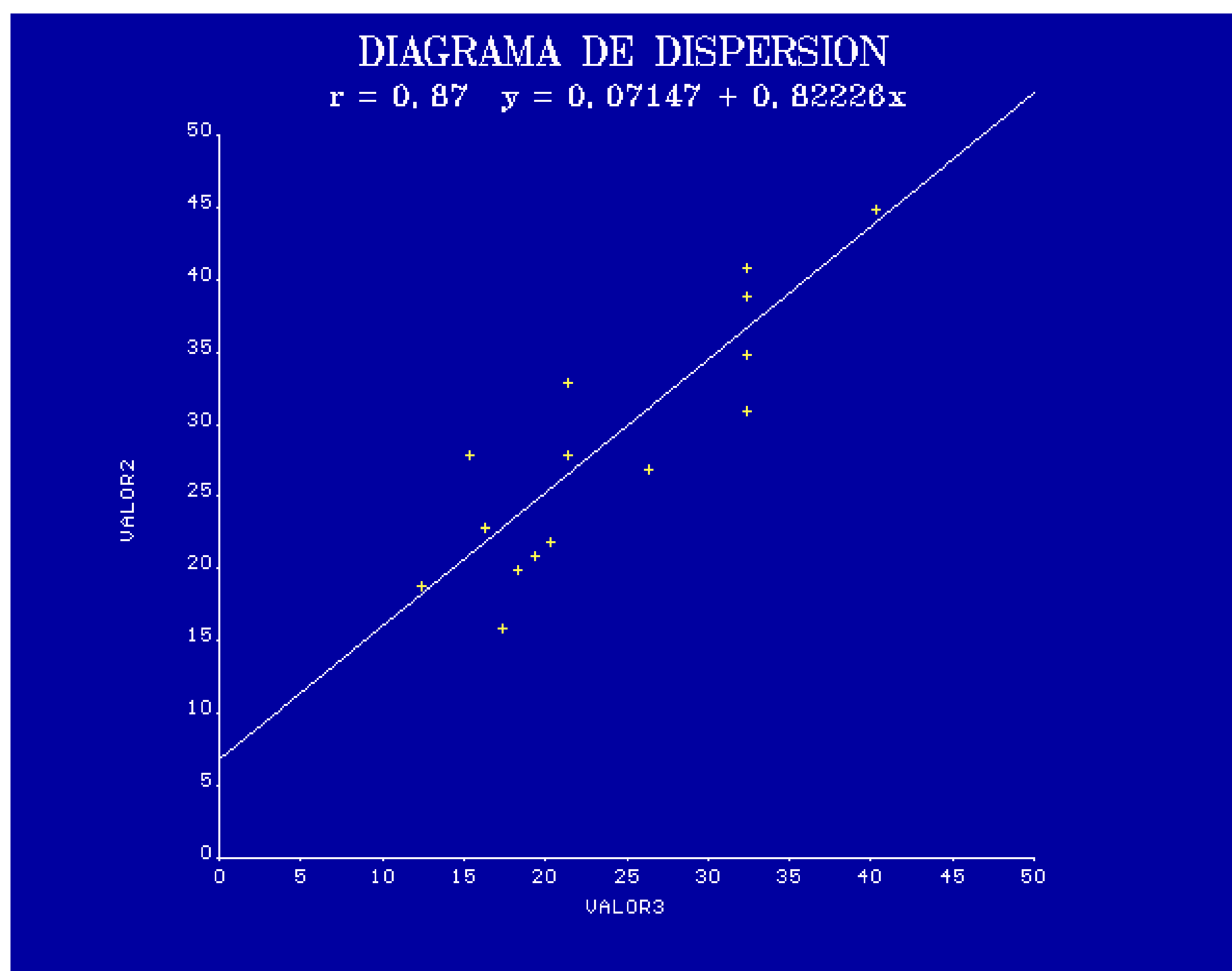
Variable	Media	Coeficiente B	Lim. Conf. al 95% Inferior	Superior	Error Est	Test-F Parcial
VALOR2	28.5333	0.8222614	0.542497	1.102026	0.129496	40.3188
Intersecc-Y		0.0714736				

Otra forma de calcular r = Suma Cuadrados Regresion / Suma cuadrados Total = 0,756

Ecuación : y=a+bx ; a = Intersecc-Y ; b = coeficiente B ; y es VALOR3 ; X es VALOR2. Por tanto y= 0,0715 + 0,8223 ó VALOR3 = 0,0715 + 0,8223*VALOR2

Valoración de r : lo que obtendríamos en la fórmula n° 14 es la raíz cuadrada de "Test F" ó E"Estadístico F" = (40,3188 =6,349, que > t(13 , 0,001) = 4,221 , por lo que se rechaza H0 a ese nivel de significación. p<0,001 . Hay una relación positiva y significativa entre Valor3 y Valor2

f) gráfico de la ecuación de regresión
(= diagrama de dispersión ="scatter")
Introduciendo hasta 5 líneas de título se puede completar el gráfico ; suele añadirse r y la ecuación :
Teclear Title 1 "\c DIAGRAMA DE DISPERSION"
Title 2 "\c r = 0,87 y = 0,0715 + 0,8223x"
SCATTER VALOR3 VALOR2 /r



EPITABLE

Trabaja con parámetros ya calculados, que vamos introduciendo cuando los pide. Son frecuencias, porcentajes, medias, varianzas, tamaños muestrales, etc. Calcula intervalos de confianza, contrasta variables, hace pruebas de conformidad, calcula números al azar, probabilidades, etc
Se pueden editar los resultados antes de imprimirlos. Para imprimir se pulsa F5. Pulsando F2 se puede abrir un fichero de texto, que luego se puede modificar en un programa de textos.

1. Proporciones o porcentajes con su IC

Teclear sucesivamente Describir, Proporción, Muestreo aleatorio simple

```
Proporci¢n, intervalo de confianza
Muestreo aleatorio Simple
Numerador                    :        7
Total de observaciones       :       12
Proporci¢n                   :  58.3333%
IC
Quadr tico de Fleiss 95% CI [28.5989-83.5010]
Binomial exacto 95% CI      [27.6670-84.8348]
Mid-p 95% CI                [30.2121-82.8309]
```

2. IC de una media
Teclear Describir , Media

```
Intervalo de confianza de una media, Alpha= 5%
Media muestral               :        10.500
Desviaci¢n est ndar muestral  :         2.200
Tama¤o muestral              :            40
Tama¤o de la poblaci¢n       :     999999999
Intervalo de confianza (95%)  :    9.82, 11.18
```

3. Comparación de porcentajes o frecuencias

<u>a) 2 muestras</u>

Teclear Comparar , Proporción , Porcentajes , 2 , OK

```
Comparaci¢n de proporciones
Muestra Porcentaje    Tama¤o muestral
---------------------------------
# 1        18.00        25
# 2        22.00        26
Un valor esperado < 5
Xý corregida de Yates     0.08
valor              :  0.776725
```

<u>b) más de 2 muestras (por ejemplo, una tabla de 2x3)</u>

Teclear Comparar, Proporción ,Tabla de datos rxc , 3 , 2 , OK

```
        5         7         9 |         21
        6         3         8 |         17
  ----------------------------------------
       11        10        17 |         38
33.3 % de los valores esperados < 5
Chiý                    1.34
Grados de libertad         2
valor                0.510797
```

4. Prueba de conformidad

Teclear Comparar, Proporción, Bondad de ajuste , 3 , OK

```
Bondad del ajuste
Clase    Observado Esperado (# o %)
-----------------------------------------
N§1           16   25.0000        25.0
N§2           28   25.0000        25.0
N§3           31   25.0000        25.0
Chi2                          5.04
Grados de libertad               2
valor                     0.080460
```

5. Contraste de medias

a) 2 muestras (t de Student)
Teclear Comparar , medias , 2 , OK

```
An lisis de la varianza
Muestra  Media       Varianza   Tama¤o muestral
-------------------------------------------
 # 1    12.00           6.00            28
 # 2    15.00           9.00            26
Varianza entre muestras       :    121.33
Varianza residual             :      7.44
Estad¡stico F                 :     16.30
valor de p                      :  0.000166
```

b) más de 2 muestras

Teclear Comparar , medias , 4 , OK

```
Análisis de la varianza
Muestra  Media       Varianza   Tamaño muestral
-------------------------------------------
 # 1    12.00           9.00            14
 # 2    13.00           8.00            18
 # 3    10.00          11.00            19
 # 4    15.00          10.00            15
Varianza entre muestras       :     73.18
Varianza residual             :      9.53
Estadístico F                 :      7.68
valor de p                        :  0.000178
```

6. Comparación de varianzas

Teclear Comparar , varianzas

```
Comparaci¢n de varianzas
Varianza N§1                 26.50
Tama¤o muestral N§1             28
Varianza N§2                 22.40
Tama¤o muestral N§2             22
F                             1.18
Valor-p de cola derecha    0.349989
Valor-p exacto 2-colas     0.699978
```

7. Estudios caso-control

Teclear: Estudios , Caso-control , No apareados

```
            Enfermos
            +       -
-----------------------------
 Caso       3      33      36
 Control   25      10      35
------------------------------
           28      43      71

Estudio de caso-control
Proporci¢n de exposici¢n
Entre casos                       10.71/100
Entre controles                   76.74/100
Test de significaci¢n
Valor-p una-cola(Fisher):          0.000000
Valor-p dos-colas(Fisher):         0.000000
Chi cuad. de Pearson Xý:29.58    p:0.000000
Chi cuad. de Yates   Xý:27.00    p:0.000000
Medidas de asociaci¢n y 95% intervalo de confiaza
Raz¢n de ventajas (OR):          0.04     0.01, 0.15
Fracci¢n prevenible             96.4%     85.4, 99.1
L¡mites de confianza exactos de la OR
Fisher:                        0.0062         0.1634
Mid-p:                         0.0079         0.1440
```

8. Eficacia vacunal

Teclear : Estudios , Método de control , Eficacia vacunal

```
Porcentaje de poblaci¢n vacunada:       78.00
Porcentaje de casos vacunados:          25.00
Eficacia vacunal                       90.60%
```

9. Valoración pruebas de cribado ("screening")

Teclear : Estudios , Cribaje

```
          Enfermedad
            +    -
------------------------------
Test +    45       3      48
Test -     5      68      73
-------------------------------
          50      71     121

Cribaje
Medidas de asociaci¢n y 95% intervalo de confiaza
Sensibilidad                  90.0%     77.4, 96.3
Especificidad                 95.8%     87.3, 98.9
Valor predictivo positivo     93.8%     81.8, 98.4
Valor predictivo negativo     93.2%     84.1, 97.5
```

10. Tamaño muestral

Teclear : Muestras , Tamaño muestral , Proporción simple

```
Tama¤o muestral, Proporci¢n simple
Tama¤o de la poblaci¢n            :    999999
Precisi¢n deseada (%)             :       5.0
Prevalencia esperada (%)          :      16.0
Efecto del Dise¤o                 :       1.0
Nivel de confianza                :       95%
Tama¤o muestral                   :       207
```

11. Números al azar (por ejemplo Primitiva)

Teclear : Muestras , Listado nº aleatorios , 6 , 1 , 49

```
 Sale        4         9        14        22        25        28
```

12. Probabilidades de una distribución binomial

Teclear Probabilidades , Dist. Binomial

Se entran los 4 datos que pide

```
Binomial: Proporci¢n vs. Estd.
Total de observaciones       :        8
Numerador                    :        4
Porcentaje esperado (%)      :    30.00
Porcentaje observado (%)     :    50.00
Probabilidad de que el # de los sucesos sea
  < 4  = 0.8058956
  <= 4 = 0.9420323
  = 4  = 0.1361367
  => 4 = 0.1941043
  > 4  = 0.0579676
Valor-p dos-colas: 0.25175236
95% intervalo de confiaza: 1-7
```

13. Probabilidades de una distribución de Poisson

Teclear : Probabilidades , Distr. Poisson

```
Poisson: Suceso raro vs. Estd.
# Observado de sucesos         3.00
# Esperado de sucesos         0.300
Probabilidad de que # de los sucesos sea
  < 3.00  = 0.9964005
  =< 3.00 = 0.9997341
  = 3.00  = 0.0033336
  => 3.00 = 0.0035994
  > 3.00 = 0.0002658
si el n£mero medio de sucesos es 0.300 ( = λ )
```

14. Prueba exacta de Fisher

Teclear : Probabilidades , Test exacto Fisher

```
    12      25      37
     9       1      10
-------------------------
    21      26      47
Test exacto de Fisher
Valor-p una-cola         :0.001544
Valor-p dos-colas        :0.002570
```

15: Permutaciones y combinaciones

Teclear : Probabilidades , Comb. Permutaciones

```
Permutaciones/Combinaciones
Número de unidades   N           49
Tomando X en el momento X         6
nº de permutaciones 10068347520
nº de combinaciones    13983816   (p.e. la Primitiva)
```

16. Probabilidades de la Distribución normal

Teclear : Probabilidades , Rango Dist. Normal

Pide la media, desviación estándar y límites del intervalo cuya p se desea calcular :

```
Rango de Distribuci¢n Normal
Media muestral                 150.00
Desviaci¢n est ndar muestral     8.00
Lower bound of range           152.00

Upper bound of range           158.00

Probabilidad de observar un valor
 < 152.00              =    0.59871
 > 152.00 y <= 158.00  =    0.24264
 > 158.00              =    0.15866
```

USO DE STATACALC

De su oferta nos resulta útil la <Tabla de 2x2> ó 2xn . Proporciona cálculos de Chi2 y sus variantes, OR, RR ,intervalos de confianza,

1. Tabla de 2x2

pide a1 , a2 , b1 y b2

```
     + Enfermo -                          An lisis de Tabla Simple
E  +--------+--------+                Odds ratio = 2.40 (0.45 <OR< 13.36)
x +|     6  |     8  | 14  L¡mites de Confianza de Cornfield (95%) para OR
p  +--------+--------+              Riesgo relativo = 1.80 (0.68 <RR< 4.77)
u -|     5  |    16  | 21  L¡mit. de Confianza (Serie de Taylor) 95% para RR
e  +--------+--------+         Ignora el R.R. es estudios de Caso-control.
s       11       24    35
t                                           Valor Chi    Valor-P
o                                         -----------   --------
                                 Sin correcci¢n :     1.41    0.2343701
                                 Mantel-Haenszel:     1.37    0.2411708
                                 Corr. de Yates :     0.67    0.4136090
                         Test exacto de Fisher: valor-P 1-cola: 0.2063255
                                                valor-P 2-colas:0.2831146

                                    Un valor esperado es menor que 5.
                                      Se recomienda test de Fisher.
                          F2 m s estratos;<Enter> No m s estratos;F10 Salir
```

pulsando **E** salen límites de confianza más exactos de la OR:

```
     + Enfermo -
E  +--------+--------+
x +|     6  |     8  | 14          ***L¡mites de Confianza Exactos***
p  +--------+--------+
u -|     5  |    16  | 21              Mehta CR, Patel NR, Gray R,
e  +--------+--------+                J. Am. Stat. Assoc.,1985,78,969-973.
s       11       24    35     Programa Pascal por ELF Franco & N Campos-Filho
t                                Ludwig Cancer Institute, Sao Paulo, Brazil
o
                                  L¡mite Exacto inferior (95%)  =  0.44
                                           Odds Ratio =  2.40
                                  L¡mite Exacto superior (95%)  = 13.22
```

2. Tabla de 2x2 con estratos

```
     + Enfermo -                          An lisis de Tabla Simple
E  +--------+--------+                Odds ratio = 0.42 (0.05 <OR< 3.01*)
x +|     5  |     6  | 11  L¡mites de Confianza de Cornfield (95%) para OR
p  +--------+--------+    *Cornfield inexacto. Usar preferentemente L¡mites
u -|     8  |     4  | 12    exactos.
e  +--------+--------+              Riesgo relativo = 0.68 (0.32 <RR< 1.46)
s      13       10    23   L¡mit. de Confianza (Serie de Taylor) 95% para RR
t                                Ignora el R.R. es estudios de Caso-control.
o
                                                 Valor Chi    Valor-P
                                                 -----------   --------
                                  Sin correcci¢n :     1.05    0.3053193
                                  Mantel-Haenszel:     1.01    0.3160728
                                  Corr. de Yates :     0.36    0.5457953
                          Test exacto de Fisher: valor-P 1-cola: 0.2734554
                                               valor-P 2-colas:0.4136492

                                     Un valor esperado es menor que 5.
                                       Se recomienda test de Fisher.

                          F2 más estratos;<Enter> No m s estratos;F10 Salir
```

se pulsa F2:

```
     + Enfermo -                    Odds ratio = 0.50 (0.06 <OR< 4.24*)
E  +--------+--------+     L¡mites de Confianza de Cornfield (95%) para OR
x +|     6  |     4  | 10 *Cornfield inexacto. Usar preferentemente L¡mites
p  +--------+--------+         exactos.
u -|     9  |     3  | 12  Riesgo relativo = 0.80 (0.44 <RR< 1.46)
e  +--------+--------+         L¡mit. de Confianza (Serie de Taylor) 95% para
RR
s      15        7    22       Ignora el R.R. es estudios de Caso-control.
t
o                                                 Valor Chi    Valor-P
                                                 -----------   --------
                                  Sin correcci¢n :     0.57    0.4519670
                                  Mantel-Haenszel:     0.54    0.4624327
                                  Corr. de Yates :     0.09    0.7699053
                           Test exacto de Fisher: valor-P 1-cola: 0.3839009
                                                valor-P 2-colas: 0.6517028

                                     Un valor esperado es menor que 5.
                                       Se recomienda test de Fisher.

                           F2 m s estratos;<Enter> No m s estratos;F10 Salir

     ***** An lisis Estratificado *****
Resumen de 2 Tablas
Odds ratio cruda para todos los estratos =  0.45
Odds Ratio Ponderada de Mantel-Haenszel=  0.45
Límites de Confianza de Cornfield 95%     0.11 <  0.45 <  1.84
Chi Resumen de Mantel-Haenszel =  0.87
Valor de P = 0.35131291
RR Crudo para todos los estratos=  0.74
Riesgo Relativo Ponderado de Mantel-Haenszel de Enfermedad,
dada la Exposici¢n=  0.74
Límites de confianza de Greenland/Robins=   0.46 < MHRR <  1.20

                                     <Enter> para otros; F10 para salir.
```

```
3. Tabla mayor de 2x2

     + Enfermo -
  +--------+--------+
E |    2   |    5   |    7
x +--------+--------+
p |    3   |    2   |    5
u +--------+--------+
e |    5   |    5   |   10
s +--------+--------+
t |    9   |    8   |   17
o +--------+--------+
      19       20       39
```

```
Análisis de Tabla Simple

    Chi =      1.52
      3 grados de libertad.
     valor p = 0.67768600


<Enter> otra tabla; F10 Salir
```

El programa **OpenStat** quiere emular al programa estrella SPSS. Es muy potente , pero está en pleno desarrollo, aún presenta algunos fallos y su manejo no es fácil.. Puede descargarse en español en
http://openstat.en.softonic.com/

y la última versión en inglés en
http://statpages.org/miller/openstat/OpenStatSetup.exe

El programa **PSPP** también emula al SPSS. Menos potente que el anterior, pero de manejo más fácil. También está en pleno desarrollo. Se puede descargar en español en
http://www.cecaps.ufmg.br/pspp/?page_id=141&lang=es

Ambos, mejor el PSPP, permiten importar los datos de un fichero de texto, incluso del más simple, como es el block de notas. Tienen su correspondiente manual.
Se verán en clase.

Tema 22. RECURSOS ESTADISTICOS EN INTERNET

Las direcciones de Internet cambian con frecuencia; las siguientes están activas en septiembre de 2008. La mayoría están en inglés, lo que no debe ser mayor inconveniente. Además ofrecen multitud de enlaces a otras páginas.

TEXTOS EN LINEA

http://www.hrc.es/bioest/estadis_1.html
(Hospital Ramón y Cajal de Madrid)

http://www.bioestadistica.uma.es/baron/apuntes
http://ftp.medprev.uma.es/libro/html.htm
(Universidad de Málaga)

http://davidmlane.com/hyperstat/

http://www.statsoft.com/textbook/stathome.html

http://faculty.vassar.edu/lowry/webtext.html

CALCULADORAS ESTADISTICAS EN LINEA

http://faculty.vassar.edu/lowry/VassarStats.html
de Richard Lowry , del Vassar College, en New York

http://www.quantitativeskills.com/sisa/index.htm
(desarrollada por el holandés Dan Uitenbroek)

http://www.physics.csbsju.edu/stats/
del College of Saint Benedict | Saint John's University, Minesota

http://statpages.org
hay programas para todo tipo de problemas estadísticos. Original de John C Pezzulo, profesor emérito de la Georgtown University de Washington.

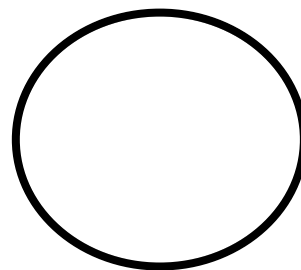
http://www.eduardobuesa.es
se puede acceder a varios programas estadísticos que resuelven la mayoría de los problemas que se tocan en esta asignatura. Se pueden descargar al propio ordenador (recomendado) o bien trabajar en línea.

Anexos

1. Formulario

2. Ejercicios

FORMULARIO DE BIOESTADISTICA



Indice

Media aritmética

a) datos originales: $\bar{X} = \frac{\sum X}{N}$ b) datos agrupados en tabla: $\bar{X} = \frac{\sum fc}{N}$

Mediana

1) datos originales (ordenados)
   a) N es impar: es el valor que ocupa el lugar (N+1)/2
   b) N es par: es la media de los valores que ocupan los lugares N/2 y siguiente

2) datos agrupados en tabla

$$M = L_i + i\left(\frac{N/2 - \sum f_M}{f_M}\right)$$

Varianza

a) datos originales: $s^2 = \frac{N\sum x^2 - (\sum x)^2}{N(N-1)}$ b) datos agrupados: $s^2 = \frac{N\sum(fc^2) - (\sum fc)^2}{N(N-1)}$

para una población el denominador se cambia a N^2

Desviación estándar

$$s = \sqrt{s^2}$$

Coeficiente de variación

$$CV = \frac{100s}{\bar{X}}$$

Percentil (Pa) : el valor que ocupa en los datos ordenados el lugar **N*a/100**

Coeficiente de correlación

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

Ecuación de regresión

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2} \quad ; \quad a = \bar{Y} - b\bar{X}$$

Probabilidad

$p = f/N$

$p(A_1$ **y** $A_2) = pA_1 * pA_2$, si son independientes; $pA_1 * p(A_2/A_1)$ si dependientes

$p(A_1$ **o** $A_2) = pA_1 + pA_2$, si incompatibles; $pA_1 + pA_2 - p(A_1 * pA_2)$, si compatibles

Distribución Binomial B(n , p)

$$\overline{X} = np \qquad s = \sqrt{npq} \qquad N = \sum N_r \qquad N_r = Np(r)$$

$$\overline{X} = \frac{\sum (rN_r)}{N} \qquad p(r) = \frac{n!}{r!*(n-r)!} p^r q^{(n-r)}$$

n	r	p 0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039

Distribución de Poisson $P(\lambda)$

$$p(r) = \frac{\lambda^r}{r!} e^{-\lambda}$$

$$\bar{X} = \lambda = np \quad ; \quad \bar{X} = \frac{\sum (r * N_r)}{N} \quad ; \quad s^2 = \lambda \quad ; \quad s = \sqrt{\lambda}$$

Valores de $e^{-\lambda}$

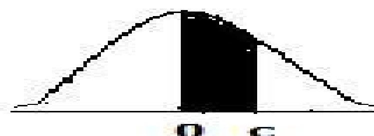
λ	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	1,0000	0,9900	0,9802	0,9704	0,9608	0,9512	0,9418	0,9324	0,9321	0,9139
0,1	0,9048	0,8958	0,8869	0,8781	0,8694	0,8607	0,8521	0,8437	0,8353	0,8270
0,2	0,8187	0,8106	0,8025	0,7945	0,7866	0,7788	0,7711	0,7634	0,7558	0,7483
0.3	0,7408	0,7334	0,7261	0,7189	0,7118	0,7047	0,6977	0,6907	0,6839	0,6771
0,4	0,6703	0,6636	0,6570	0,6505	0,6440	0,6376	0,6313	0,6250	0,6188	0,6126
0,5	0,6065	0,6005	0,5945	0.5886	0,5827	0,5770	0,5712	0,5655	0,5599	0,5543
0.6	0,5488	0,5434	0,5379	0,5326	0,5273	0,5220	0,5169	0,5117	0,5066	0,5016
0.7	0,4966	0,4916	0,4868	0.4819	0,4771	0,4724	0,4677	0,4630	0,4584	0,4538
0.8	0,4493	0,4449	0,4404	0,4360	0,4317	0,4274	0,4232	0,4190	0,4148	0,4107
0,9	0,4066	0,4025	0,3985	0,3946	0,3906	0,3867	0,3829	0,3791	0,3753	0,3716
λ	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	0,36788	0,13534	0,04979	0,01832	0,006738	0,002479	0,000912	0,000335	0,000123	0,000045

Para obtener otros valores de $e^{-\lambda}$ recordar las normas del producto de potencias y manejar adecuadamente los exponentes.

Así: $e^{-1,25} = e^{-1} * e^{-0,25} = 0,36788*0,7788 = 0,2865$

Distribución normal N (0 , 1)

$$c = \frac{x - \bar{x}}{s}$$



la tabla da la probabilidad de que un valor cualquiera esté entre c = 0 y otro valor de c

a esta c se la llama hoy día mayoritariamente **z**

c	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	**0,1915**	**0,1950**	**0,1985**	**0,2019**	**0,2054**	**0,2088**	**0,2123**	**0,2157**	**0,2190**	**0,2224**
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	**0,3413**	**0,3438**	**0,3461**	**0,3485**	**0,3508**	**0,3531**	**0,3554**	**0,3577**	**0,3599**	**0,3621**
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	**0,4332**	**0,4345**	**0,4357**	**0,4370**	**0,4382**	**0,4394**	**0,4406**	**0,4418**	**0,4429**	**0,4441**
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	**0,4772**	**0,4778**	**0,4783**	**0,4788**	**0,4793**	**0,4798**	**0,4803**	**0,4808**	**0,4812**	**0,4817**
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	**0,4938**	**0,4940**	**0,4941**	**0,4943**	**0,4945**	**0,4946**	**0,4948**	**0,4949**	**0,4951**	**0,4952**
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	**0,4987**	**0,4987**	**0,4987**	**0,4988**	**0,4988**	**0,4989**	**0,4989**	**0,4989**	**0,4990**	**0,4990**
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**	**0,4998**
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Tabla de $\chi 2$

Probabilidad

g. l.	**0,05**	**0,02**	**0,01**	**0,001**
1	3,84	5,41	6,64	10,83
2	5,99	7,82	9,21	13,82
3	7,81	9,84	11,34	16,27
4	9,49	11,77	13,28	18,47
5	11,07	13,39	15,09	20,52
6	12,59	15,03	16,81	22,46
7	14,07	16,62	18,48	24,32
8	15,51	18,17	20,09	26,13
9	16,92	19,68	21,67	27,88
10	18,31	21,16	23,21	29,59

Tabla de la t de Student

g. l.	**0,05**	**0,01**	**0,001**		**g. l.**	**0,05**	**0,01**	**0,001**
1	12,71	63,66	636,6		**26**	2,056	2,779	3,707
2	4,303	9,925	31,60		**27**	2,052	2,771	3,690
3	3,182	5,841	12,94		**28**	2,048	2,763	3,674
4	2,776	4,604	8,610		**29**	2,045	2,756	3,659
5	2,571	4,032	6,859		**30**	2,042	2,750	3,646
6	2,447	3,707	5,959		**35**	2,030	2,724	3,592
7	2,365	3,499	5,405		**40**	2,021	2,704	3,551
8	2,306	3,355	5,041		**45**	2,014	2,689	3,521
9	2,262	3,250	4,781		**50**	2,008	2,678	3,496
10	2,228	3,169	4,587		**55**	2,004	2,669	3,476
11	2,201	3,106	4,437		**60**	2,000	2,660	3,460
12	2,179	3,055	4,318		**70**	1,994	2,648	3,435
13	2,160	3,012	4,221		**80**	1,989	2,638	3,416
14	2,145	2,977	4,140		**90**	1,986	2,631	3,402
15	2,131	2,947	4,073		**100**	1,982	2,626	3,390
16	2,120	2,921	4,015		**120**	1,980	2,617	3,373
17	2,110	2,898	3,965		**130**	1,977	2,612	3,361
18	2,101	2,878	3,922		**140**	1,975	2,607	3,352
19	2,093	2,861	3,883		**150**	1,974	2,605	3,349
20	2,086	2,845	3,850		**160**	1,973	2,603	3,346
21	2,080	2,831	3,819		**200**	1,972	2,601	3,340
22	2,074	2,819	3,792		**300**	1,968	2,592	3,340
23	2,069	2,807	3,767		**400**	1,966	2,588	3,315
24	2,064	2,797	3,745		**500**	1,965	2,586	3,310
25	2,060	2,787	3,725		∞	1,960	2,576	3,291

F de Snedecor-Fisher $\alpha = 0,05$

g.l. 1 →

g.l. 2 ↓	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500	∞
5	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,56	4,52	4,50	4,48	4,46	4,44	4,41	4,37	4,36
6	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,87	3,83	3,81	3,79	3,77	3,75	3,71	3,68	3,67
7	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51	3,44	3,40	3,38	3,36	3,34	3,32	3,27	3,24	3,23
8	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22	3,15	3,11	3,08	3,06	3,04	3,02	2,97	2,94	2,93
9	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01	2,94	2,89	2,86	2,84	2,83	2,80	2,76	2,72	2,71
10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85	2,77	2,73	2,70	2,68	2,66	2,64	2,59	2,55	2,54
11	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72	2,65	2,60	2,57	2,55	2,53	2,51	2,46	2,42	2,40
12	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62	2,54	2,50	2,47	2,44	2,43	2,40	2,35	2,31	2,30
13	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53	2,46	2,41	2,38	2,36	2,34	2,31	2,26	2,22	2,21
14	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46	2,39	2,34	2,31	2,28	2,27	2,24	2,19	2,14	2,13
15	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40	2,33	2,28	2,25	2,22	2,20	2,18	2,12	2,08	2,07
16	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35	2,28	2,23	2,19	2,17	2,15	2,12	2,07	2,02	2,01
17	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31	2,23	2,18	2,15	2,12	2,10	2,08	2,02	1,97	1,96
18	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27	2,19	2,14	2,11	2,08	2,06	2,04	1,98	1,93	1,92
19	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,16	2,11	2,07	2,05	2,03	2,00	1,94	1,89	1,88
20	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20	2,12	2,07	2,04	2,01	1,99	1,97	1,91	1,86	1,84
21	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,22	2,20	2,18	2,10	2,05	2,01	1,98	1,96	1,94	1,88	1,83	1,81
22	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,20	2,17	2,15	2,07	2,02	1,98	1,96	1,94	1,91	1,85	1,80	1,78
23	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,18	2,15	2,13	2,05	2,00	1,96	1,93	1,91	1,88	1,82	1,77	1,76
24	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,15	2,13	2,11	2,03	1,97	1,94	1,91	1,89	1,86	1,80	1,75	1,73
25	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,11	2,09	2,01	1,96	1,92	1,89	1,87	1,84	1,78	1,73	1,71
26	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,12	2,09	2,07	1,99	1,94	1,90	1,87	1,85	1,82	1,76	1,71	1,69
27	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13	2,10	2,08	2,06	1,97	1,92	1,88	1,86	1,84	1,81	1,74	1,69	1,67
28	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,09	2,06	2,04	1,96	1,91	1,87	1,84	1,82	1,79	1,73	1,67	1,65
29	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,08	2,05	2,03	1,94	1,89	1,85	1,83	1,81	1,77	1,71	1,65	1,64
30	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,06	2,04	2,01	1,93	1,88	1,84	1,81	1,79	1,76	1,70	1,64	1,62
40	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,97	1,95	1,92	1,84	1,78	1,74	1,72	1,69	1,66	1,59	1,53	1,51
50	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99	1,95	1,92	1,89	1,87	1,78	1,73	1,69	1,66	1,63	1,60	1,52	1,46	1,44
75	2,73	2,49	2,34	2,22	2,13	2,06	2,01	1,96	1,92	1,88	1,85	1,83	1,80	1,71	1,65	1,61	1,58	1,55	1,52	1,44	1,36	1,34
100	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,89	1,85	1,82	1,79	1,77	1,68	1,62	1,57	1,54	1,52	1,48	1,39	1,31	1,29
200	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,84	1,80	1,77	1,74	1,72	1,62	1,56	1,52	1,48	1,46	1,41	1,32	1,22	1.19
1000	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,80	1,76	1,73	1,70	1,68	1,58	1,52	1,47	1,43	1,41	1,36	1,26	1,13	1,08
∞	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,72	1,69	1,67	1,57	1,51	1,46	1,42	1,40	1,35	1,24	1,11	1

F de Snedecor-Fisher $\alpha = 0,01$

g.l. 1 →

g.l. 2 ↓	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500	∞
5	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,82	9,77	9,72	9,55	9,45	9,38	9,33	9,29	9,24	9,13	9,04	9,02
6	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,40	7,30	7,23	7,18	7,14	7,09	6,99	6,90	6,88
7	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,36	6,31	6,16	6,06	5,99	5,94	5,91	5,86	5,75	5,67	5,65
8	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,61	5,56	5,52	5,36	5,26	5,20	5,15	5,12	5,07	4,96	4,88	4,86
9	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,01	4,96	4,81	4,71	4,65	4,60	4,57	4,52	4,41	4,33	4,31
10	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,65	4,60	4,56	4,41	4,31	4,25	4,20	4,17	4,12	4,01	3,93	3,91
11	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,10	4,01	3,94	3,89	3,86	3,81	3,71	3,62	3,60
12	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,86	3,76	3,70	3,65	3,62	3,57	3,47	3,38	3,36
13	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,91	3,86	3,82	3,66	3,57	3,51	3,46	3,43	3,38	3,27	3,19	3,17
14	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,51	3,41	3,35	3,30	3,27	3,22	3,11	3,03	3,00
15	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,37	3,28	3,21	3,17	3,13	3,08	2,98	2,89	2,87
16	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,50	3,45	3,41	3,26	3,16	3,10	3,05	3,02	2,97	2,86	2,78	2,75
17	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,16	3,07	3,00	2,96	2,92	2,87	2,76	2,68	2,65
18	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,32	3,27	3,23	3,08	2,98	2,92	2,87	2,84	2,78	2,68	2,59	2,57
19	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,00	2,91	2,84	2,80	2,76	2,71	2,60	2,51	2,49
20	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,18	3,13	3,09	2,94	2,84	2,78	2,73	2,69	2,64	2,54	2,44	2,42
21	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,12	3,07	3,03	2,88	2,79	2,72	2,67	2,64	2,58	2,48	2,38	2,36
22	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	3,07	3,02	2,98	2,83	2,73	2,67	2,62	2,58	2,53	2,42	2,33	2,31
23	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	3,02	2,97	2,93	2,78	2,69	2,62	2,57	2,54	2,48	2,37	2,28	2,26
24	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,98	2,93	2,89	2,74	2,64	2,58	2,53	2,49	2,44	2,33	2,24	2,21
25	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,94	2,89	2,85	2,70	2,60	2,54	2,49	2,45	2,40	2,29	2,19	2,17
26	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,90	2,86	2,81	2,66	2,57	2,50	2,45	2,42	2,36	2,25	2,16	2,13
27	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99	2,93	2,87	2,82	2,78	2,63	2,54	2,47	2,42	2,38	2,33	2,22	2,12	2,10
28	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,96	2,90	2,84	2,79	2,75	2,60	2,51	2,44	2,39	2,35	2,30	2,19	2,09	2,08
29	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93	2,87	2,81	2,77	2,73	2,57	2,48	2,41	2,36	2,33	2,27	2,16	2,06	2,03
30	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,79	2,74	2,70	2,55	2,45	2,39	2,34	2,30	2,25	2,13	2,03	2,01
40	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,61	2,56	2,52	2,37	2,27	2,20	2,15	2,11	2,06	1,94	1,83	1,80
50	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,63	2,56	2,51	2,46	2,42	2,27	2,17	2,10	2,05	2,01	1,95	1,82	1,71	1,68
75	4,05	3,58	3,27	3,05	2,89	2,76	2,65	2,57	2,49	2,43	2,38	2,33	2,29	2,13	2,03	1,96	1,91	1,87	1,81	1,67	1,55	1,52
100	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,43	2,37	2,31	2,27	2,22	2,07	1,97	1,89	1,84	1,80	1,74	1,60	1,47	1,43
200	3,88	3,41	3,11	2,89	2,73	2,60	2,50	2,41	2,34	2,27	2,22	2,17	2,13	1,97	1,87	1,79	1,74	1,69	1,63	1,48	1,33	1,29
1000	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,27	2,20	2,15	2,10	2,06	1,90	1,79	1,72	1,66	1,61	1,54	1,38	1,19	1,11
∞	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,25	2,18	2,13	2,08	2,04	1,88	1,77	1,70	1,64	1,59	1,52	1,36	1,15	1

F de Snedecor-Fisher $\alpha = 0{,}001$

g.l. 1 →

g.l. 2 ↓	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	35	40	50	100	500	∞
5	33,2	31,1	29,8	28,8	28,2	27,6	27,2	26,9	26,6	26,4	26,2	26,1	25,9	25,4	25,1	24,9	24,7	24,6	24,4	24,1	23,9	23,79
6	23,7	21,9	20,8	20,0	19,5	19,0	18,7	18,4	18,2	18,0	17,8	17,7	17,6	17,1	16,9	16,7	16,5	16,4	16,3	16,0	15,8	15,75
7	18,8	17,2	16,2	15,5	15,0	14,6	14,3	14,1	13,9	13,7	13,6	13,4	13,3	12,9	12,7	12,5	12,4	12,3	12,2	12,0	11,7	11,70
8	15,8	14,4	13,5	12,9	12,4	12,0	11,8	11,5	11,4	11,2	11,1	10,9	10,8	10,5	10,3	10,1	10,0	9,9	9,8	9,6	9,4	9,33
9	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89	9,72	9,57	9,44	9,33	9,24	8,90	8,69	8,55	8,45	8,37	8,26	8,04	7,86	7,81
10	12,55	11,28	10,48	9,93	9,52	9,20	8,96	8,75	8,59	8,45	8,32	8,22	8,13	7,80	7,60	7,47	7,37	7,30	7,19	6,98	6,81	6,76
11	11,56	10,35	9,58	9,05	8,65	8,35	8,12	7,92	7,76	7,63	7,51	7,41	7,32	7,01	6,81	6,68	6,59	6,52	6,42	6,21	6,04	6,00
12	10,80	9,63	8,89	8,38	8,00	7,71	7,48	7,29	7,14	7,00	6,89	6,79	6,71	6,40	6,22	6,09	6,00	5,93	5,83	5,63	5,46	5,42
13	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80	6,65	6,52	6,41	6,31	6,23	5,93	5,75	5,63	5,54	5,47	5,37	5,17	5,01	4,97
14	9,73	8,62	7,92	7,44	7,08	6,80	6,58	6,40	6,26	6,13	6,02	5,93	5,85	5,56	5,38	5,25	5,17	5,10	5,00	4,81	4,65	4,60
15	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08	5,94	5,81	5,71	5,62	5,54	5,25	5,07	4,95	4,86	4,80	4,70	4,51	4,35	4,31
16	9,01	7,94	7,27	6,80	6,46	6,20	5,98	5,81	5,67	5,55	5,44	5,35	5,27	4,99	4,82	4,70	4,61	4,54	4,45	4,26	4,10	4,06
17	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58	5,44	5,32	5,22	5,13	5,05	4,78	4,60	4,48	4,40	4,33	4,24	4,05	3,89	3,85
18	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39	5,25	5,13	5,03	4,94	4,87	4,59	4,42	4,30	4,22	4,15	4,06	3,87	3,71	3,67
19	8,28	7,27	6,62	6,18	5,85	5,59	5,39	5,22	5,08	4,97	4,87	4,78	4,70	4,43	4,26	4,14	4,06	3,99	3,90	3,71	3,55	3,51
20	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08	4,94	4,82	4,72	4,64	4,56	4,29	4,12	4,00	3,92	3,86	3,77	3,58	3,42	3,38
21	7,94	6,95	6,32	5,88	5,56	5,31	5,11	4,95	4,81	4,70	4,60	4,51	4,44	4,17	4,00	3,88	3,80	3,74	3,64	3,46	3,30	3,26
22	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83	4,70	4,58	4,49	4,40	4,33	4,06	3,89	3,78	3,69	3,63	3,54	3,35	3,19	3,15
23	7,67	6,70	6,08	5,65	5,33	5,09	4,89	4,73	4,60	4,48	4,39	4,30	4,23	3,96	3,79	3,68	3,60	3,53	3,44	3,25	3,10	3,05
24	7,55	6,59	5,98	5,55	5,24	4,99	4,80	4,64	4,51	4,39	4,30	4,21	4,14	3,87	3,71	3,59	3,51	3,45	3,36	3,17	3,01	2,97
25	7,45	6,49	5,89	5,46	5,15	4,91	4,71	4,56	4,42	4,31	4,22	4,13	4,06	3,79	3,63	3,52	3,43	3,37	3,28	3,09	2,93	2,89
26	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48	4,35	4,24	4,14	4,06	3,99	3,72	3,56	3,44	3,36	3,30	3,21	3,02	2,86	2,82
27	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41	4,28	4,17	4,08	3,99	3,92	3,66	3,49	3,38	3,30	3,23	3,14	2,96	2,80	2,75
28	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35	4,22	4,11	4,01	3,93	3,86	3,60	3,43	3,32	3,24	3,18	3,09	2,90	2,74	2,69
29	7,12	6,19	5,59	5,18	4,87	4,64	4,45	4,29	4,16	4,05	3,96	3,88	3,80	3,54	3,38	3,27	3,18	3,12	3,03	2,84	2,68	2,64
30	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24	4,11	4,00	3,91	3,82	3,75	3,49	3,33	3,22	3,13	3,07	2,98	2,79	2,63	2,59
40	6,59	5,70	5,13	4,73	4,44	4,21	4,02	3,87	3,75	3,64	3,55	3,47	3,40	3,15	2,98	2,87	2,79	2,73	2,64	2,44	2,28	2,23
50	6,34	5,46	4,90	4,51	4,22	4,00	3,82	3,67	3,55	3,44	3,35	3,27	3,20	2,95	2,79	2,68	2,60	2,53	2,44	2,25	2,07	2,03
75	6,01	5,16	4,62	4,24	3,96	3,74	3,56	3,42	3,30	3,19	3,10	3,03	2,96	2,71	2,55	2,44	2,35	2,29	2,19	1,99	1,81	1,75
100	5,86	5,02	4,48	4,11	3,83	3,61	3,44	3,30	3,18	3,07	2,99	2,91	2,84	2,59	2,43	2,32	2,24	2,17	2,08	1,87	1,67	1,62
200	5,63	4,81	4,29	3,92	3,65	3,43	3,26	3,12	3,00	2,90	2,82	2,74	2,67	2,42	2,26	2,15	2,07	2,00	1,90	1,68	1,46	1,39
1000	5,46	4,65	4,14	3,78	3,51	3,30	3,13	2,99	2,87	2,77	2,69	2,61	2,54	2,30	2,14	2,02	1,94	1,87	1,77	1,53	1,27	1,15
∞	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96	2,84	2,74	2,66	2,58	2,51	2,27	2,10	1,99	1,90	1,84	1,73	1,49	1,21	1

Gráfico de David

Intervalos del 95 por 100 para el coeficiente de correlación *

(De F. N. DAVID, *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, The Biometrika Office. Londres, 1938.)

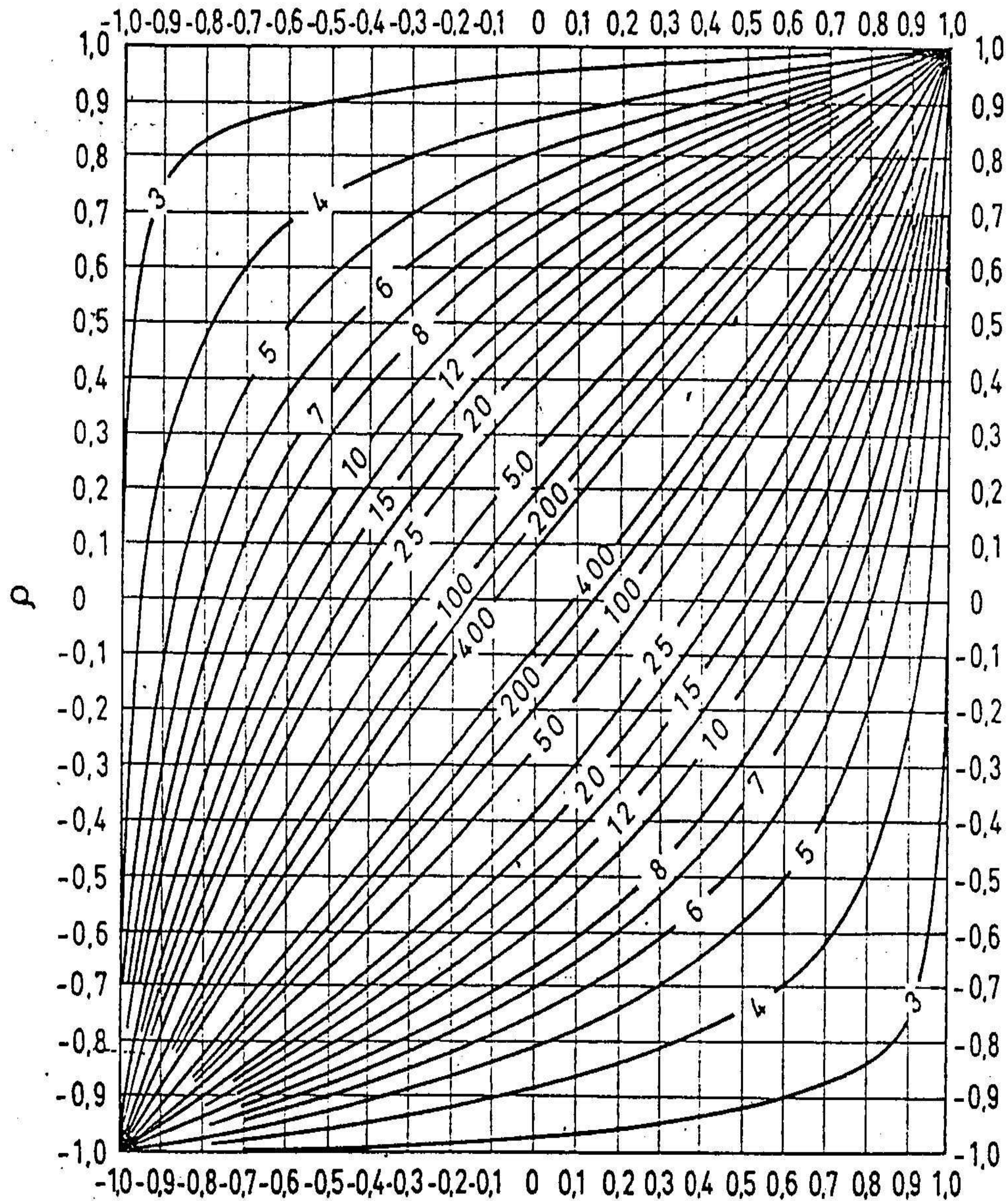


Tabla de Zr

Valores de r para distintos valores de z_r

(De *Statistical Methods for Research Workers*, por R. A. FISCHER, Oliver and Boyd. Edimburgo.)

z_r	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0699	0,0798	0,0898
0,1	0,0997	0,1096	0,1194	0,1293	0,1391	0,1489	0,1587	0,1684	0,1781	0,1878
0,2	0,1974	0,2070	0,2165	0,2260	0,2355	0,2449	0,2543	0,2636	0,2729	0,2821
0,3	0,2913	0,3004	0,3095	0,3185	0,3275	0,3364	0,3452	0,3540	0,3627	0,3714
0,4	0,3800	0,3885	0,3969	0,4053	0,4136	0,4219	0,4301	0,4382	0,4462	0,4542
0,5	0,4621	0,4700	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299
0,6	0,5370	0,5441	0,5511	0,5581	0,5649	0,5717	0,5784	0,5850	0,5915	0,5980
0,7	0,6044	0,6107	0,6169	0,6231	0,6291	0,6352	0,6411	0,6469	0,6527	0,6584
0,8	0,6640	0,6696	0,6751	0,6805	0,6858	0,6911	0,6963	0,7014	0,7064	0,7114
0,9	0,7163	0,7211	0,7259	0,7306	0,7352	0,7398	0,7443	0,7487	0,7531	0,7574
1,0	0,7616	0,7658	0,7699	0,7739	0,7779	0,7818	0,7857	0,7895	0,7932	0,7969
1,1	0,8005	0,8041	0,8076	0,8110	0,8144	0,8178	0,8210	0,8243	0,8375	0,8306
1,2	0,8337	0,8367	0,8397	0,8426	0,8455	0,8483	0,8511	0,8538	0,8565	0,8591
1,3	0,8617	0,8643	0,8668	0,8693	0,8717	0,8741	0,8764	0,8787	0,8810	0,8832
1,4	0,8854	0,8875	0,8896	0,8917	0,8937	0,8957	0,8977	0,8996	0,9015	0,9033
1,5	0,9052	0,9069	0,9087	0,9104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9202
1,6	0,9217	0,9232	0,9246	0,9261	0,9275	0,9289	0,9302	0,9316	0,9329	0,9342
1,8	0,9354	0,9367	0,9379	0,9391	0,9402	0,9414	0,9425	0,9436	0,9447	0,9458
1,7	0,9468	0,9478	0,9498	0,9488	0,9508	0,9518	0,9527	0,9536	0,9545	0,9554
1,9	0,9562	0,9571	0,9579	0,9587	0,9595	0,9603	0,9611	0,9619	0,9626	0,9633
2,0	0,9640	0,9647	0,9654	0,9661	0,9668	0,9674	0,9680	0,9687	0,9693	0,9699
2,1	0,9705	0,9710	0,9716	0,9722	0,9727	0,9732	0,9738	0,9743	0,9748	0,9753
2,2	0,9757	0,9762	0,9767	0,9771	0,9776	0,9780	0,9785	0,9789	0,9793	0,9797
2,3	0,9801	0,9805	0,9809	0,9812	0,9816	0,9820	0,9823	0,9827	0,9830	0,9834
2,4	0,9837	0,9840	0,9843	0,9846	0,9849	0,9852	0,9855	0,9858	0,9861	0,9863
2,5	0,9866	0,9869	0,9871	0,9874	0,9876	0,9879	0,9881	0,9884	0,9886	0,9888
2,6	0,9890	0,9892	0,9895	0,9897	0,9899	0,9901	0,9903	0,9905	0,9906	0,9908
2,7	0,9910	0,9912	0,9914	0,9915	0,9917	0,9919	0,9920	0,9922	0,9923	0,9925
2,8	0,9926	0,9928	0,9929	0,9931	0,9932	0,9933	0,9935	0,9936	0,9937	0,9938
2,9	0.9940	0,9941	0,9942	0,9943	0,9944	0,9945	0,9946	0,9947	0,9949	0,9950
3,0	0,9951									
4,0	0.9993									
5,0	0,9999									

*** **Valores de referencia de c en la DN**

α ó p	Prueba bilateral	Prueba unilateral
0,05	**1,96**	1,64
0,01	**2,58**	2,33
0,001	**3,30**	3,09

*** Tamaño muestral

1) para estimación de población infinita

de una media : $N = \left(\frac{c * s}{k} \right)^2$; de p ó % : $N = \frac{c^2 pq}{k^2}$

2) para contraste de variables (N por muestra)

- de medias : $N = 13 * s^2 / d^2$
- de 2 proporciones o porcentajes : $N = 6'5(p_1q_1+p_2q_2)/d^2$

*** Estimación de parámetros . Intervalo de confianza

1) **de una media**

muestra grande $I_{\bar{X}} = \bar{X} \pm c \frac{s}{\sqrt{N}}$; muestra pequeña $I_{\bar{X}} = \bar{X} \pm t_{(n-1)} \frac{s}{\sqrt{N}}$

2) **de un porcentaje o proporción**

muestra grande $I_p = p \pm c \sqrt{\frac{pq}{N}}$; muestra pequeña $I_p = p \pm t_{(n-1)} \sqrt{\frac{pq}{N}}$

3) **de un coeficiente de correlación r**

a. cálculo de Zr : $\mathbf{Zr} = \left(\frac{1}{2} \ln \frac{1+r}{1-r} \right) \pm \frac{c}{\sqrt{N-3}}$ y ver la tabla de Zr

o bien a partir de la Zr obtenida:

b. invertir Zr : para cada Zr $\mathbf{r} := \frac{e^{2Zr} - 1}{e^{2Zr} + 1}$

*** Pruebas de conformidad

1) **proporción o porcentaje** $Z = \frac{p_m - p_p}{\sqrt{\frac{p_p * q_p}{N}}}$

Valoración : ♦ *muestra grande : por DN ;* ♦ *muestra pequeña: se multiplican **p** y **q** de la población por N ---si ambos productos son ≥ 5 (ó 500, si es %): por DN; ---si alguno de ellos es < 5 (ó 500): por* t_{n-1}

2) **frecuencias:** usar fórmula de contraste nº 3 *Valorar por* χ^2 *,g.l. = nº de modalidades – 1*

3) media

$\mathbf{Z} = \frac{(\bar{X}_m - \bar{X}_p)\sqrt{N}}{s}$ *Valorar por DN, si es muestra grande; si es pequeña por* t_{n-1}

***Pruebas de contraste de hipótesis

Variables	D a t o s	Modalidades	C o n t r a s t e d e . . .	F ó r m u l a n º
CL y CL	Independ.	2	**Proporciones o porcentajes**	1
		2	**Frecuencias**	2
		3 ó más	**Frecuencias**	3
	Aparead.	2	Prueba de comparación ■ **proporciones ó %** ■ **frecuencias** ------------------------- prueba de relación ■ **proporciones ó %** ■ **frecuencias**	 4 5 ---------------------- 1 2
CL y CT	Independ.	2	**dos medias** ■ paramétrico ■ no paramétrico	 6 7 Mann-Whitney
		3 ó más	**k medias** ■ paramétrico ■ no paramétrico	 8 ANOVA-1 9 Kruskal-Wallis
	Aparead.	2	**2 medias** - prueba de comparación ■ paramétrico ■ no paramétrico ------------------------- - prueba de relación como si fuera CT y CT	 10 11 Pª de los signos ---------------------- 14 ó 15
		3 ó más	**k medias** (prueba de comparación) ■ paramétrico ■ no paramétrico	 12 ANOVA-2 13 Test de Friedman
CT y CT	Todos	2	**Coeficiente de correlación** ■ paramétrico ■ no paramétrico	 14 (r de Pearson) 15 (r de Spearman)

(es más fácil utilizar porcentajes que proporciones)

Fórmula nº 1

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_0 q_0}{N_1} + \frac{p_0 q_0}{N_2}}} \quad \text{siendo } \mathbf{p_0} = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$

Valoración: si N1 y N2 *≥30 , por la DN*

si N1 ó N2 <30

a) si p_0N_1 , q_0N_1, p_0N_2 y q_0N_2 ≥5 (ó 500 si es %) por DN

b) *si algún producto <5 pero > 3 : por t (N_1+N_2 –2)*

c) si algún producto < 3 : por p exacta de Fisher

Fórmula nº 2

$$Z = \frac{N\left(a_1 b_2 - a_2 b_1\right)^2}{N_a N_b N_1 N_2}$$

tabla:

a_1	a_2	N_a
b_1	b_2	N_b
N_1	N_2	N

Condición de aplicación: todas las E ≥ 5 ; Valoración: por χ^2_1

Si **alguna E<5 , pero ≥ 3: usar fórmula de Yates**

Si alguna E<3: calcular p exacta de Fisher

Fórmula de Yates:
$$Z = \frac{N\left(\left|a_1 b_2 - a_2 b_1\right| - \frac{N}{2}\right)^2}{N_a N_b N_1 N_2}$$

*** **p exacta de Fisher**
$$p_F = \sum_{a_1 = a_1}^{a_1 = 0} \frac{N_1!\,N_2!\,N_a!\,N_b!}{a_1!\,b_1!\,a_2!\,b_2!\,N!}$$

Fórmula nº 3

- **Si todas las E ≥ 5 :** $Z = \sum \frac{(O - E)^2}{E}$
- **Si alguna E<5 pero ≥ 3:** $Z = \sum \frac{(\left|O - E\right| - 0,5)^2}{E}$

Valoración: por $\chi^2_{(f-1)(k-1)}$

Fórmula nº 4 : contraste de proporciones (si se utilizan % hay que dividir por 100)

$Z = (p_1 - 0,5)\sqrt{4N}$, ***siendo*** $N = N_1 + N_2$

$N_1 = n^o$ de A+ B- ; $N_2 = n^o$ de A- B+ ; $p_1 = \frac{N_1}{N}$

¡esta N no es la N de la tabla!

Valoración :

si N ≥ 10 por DN ; si <10 pero ≥ 5 por t_{N-1}; si <5: p exacta de Fisher

Fórmula nº 5 : contraste de frecuencias (más sencilla que la anterior)

los símbolos son los mismos de la fórmula nº 4

Si N ≥ 10: $Z = \frac{(N_1 - N_2)^2}{N}$ **; Si N<10 y ≥ 5 :** $Z = \frac{(\left|N_1 - N_2\right| - 1)^2}{N}$

Valoración **: por χ^2_1** ***si*** **N<5 : p exacta de Fisher**

Fórmula nº 6

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}} \text{ , siendo } s^2 = \frac{s_1^2(N_1 - 1) + s_2^2(N_2 - 1)}{N_1 + N_2 - 2}$$

(s^2 es la varianza común)

Condición para muestras pequeñas: V< F(N_1-1 , N_2-1 , 0,05) ; V= s^2 mayor/ s^2 menor; este N_1 es el que corresponde a la varianza mayor ; N_2 el de la menor
Valoración: si ambas muestras son grandes por c de la DN , si alguna es pequeña por $t_{N1+N2-2}$

Fórmula nº 7 Prueba de Mann-Whitney

Se ordenan todos los datos a la vez

Individ.		X_1	R		X_2	R
1						
2						
3						
4						
5						
6						
.....						
n						
		Suma			Suma	
			R_1			R_2

$N = n_1 + n_2$

$$Z = \frac{R - \frac{n(N+1)}{2}}{\sqrt{\frac{n_1 n_2 (N+1)}{12}}}$$

R es la menor de R1 y R2 ;
n es el tamaño de la muestra que corresponde a esa R

Prueba de que se ha calculado bien : $R_1 + R_2 = N(N+1)/2$

Valoración: si ambas muestras grandes por c de la DN, si alguna es pequeña por t_{N-2}

Fórmula nº 8 ANOVA-1

a) a partir de los datos originales: utilizar la plantilla de la página siguiente

b) a partir de la media, varianza y el tamaño n de cada muestra se puede calcular Σx y ΣX^2 luego se pasa el resultado a la plantilla:

$$\sum X = n\bar{X}$$

$$\sum X^2 = \frac{s^2 n(n-1) + \left(\sum X\right)^2}{n}$$

ANOVA-1

Muestras →	1		2		3		……		k		
Individuos ↓	X	X^2	X	X^2	X	X^2	X	X^2	X	X^2	Valoración: por F(k-1 , N-k , α)
1											
2											
3											
4											
5											
6											Si se rechaza H_0 hay que aplicar prueba de Scheffé, de dos en dos, ordenados por su media
………											
ΣX											ΣΣX = B
$(\Sigma X)^2$											
n											Σn = N
$(\Sigma X)^2/n$											$\Sigma[(\Sigma X)^2/n] = A$
ΣX^2											$\Sigma(\Sigma X^2) = C$
$\bar{X}$											

$$C_A = A - \frac{B^2}{N} =$$

$$V_A = \frac{C_A}{k-1} =$$

$$C_T = C - \frac{B^2}{N} =$$

$$Z = \frac{V_A}{V_R} =$$

$$V_R = \frac{C_R}{N-k} =$$

$$C_R = C_T - C_A =$$

Fórmula nº 8 bis (Scheffé)

Pasos

1) ordenar las medias, de mayor a menor o viceversa

2 compararlas por parejas, empezando por las de los extremos

3) aplicar la fórmula 8 bis

$$Z_{sch} = \frac{(\bar{X}_i - \bar{X}_j)^2}{V_R (k-1)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Valoración por $F_{k-1, N-k}$

Fórmula nº 9 PRUEBA DE KRUSKAL-WALLIS

Individ.	Muestras									
	1		2		3		....		k	
	x_1	R	x_2	R	x_3	R	x.	R	x_k	R
1										
2										
3										
4										
5										
6										
7										
8										
…										
n										

***** Se ordenan los datos de todas las muestras en conjunto.**

***** Valoración por $\chi^2_{(k-1\,,\,\alpha)}$**

***** Si se rechaza H_0 hay que aplicar la prueba de Mann-Whitney de dos en dos, ordenados por su T**

							Suma
$T = \Sigma R$							
T^2							
n						→ $N = \Sigma n$ →	
T^2/n						→ $\Sigma(T^2/n)$ →	

$$Z = \left[\frac{12}{N(N+1)}\sum\left(\frac{T^2}{n}\right)\right] - 3(N+1)$$

Fórmula nº 10

$$Z = \bar{X}_d\sqrt{\frac{N}{s_d^2}}$$

Valoración: muestra grande por c de la DN; si pequeña por t_{N-1}

Fórmula nº 11 Test de los signos

$$Z = \frac{(2x-N)}{\sqrt{N}}$$, **siendo x el mayor de N_1 y N_2 y $N = N_1 + N_2$**

***Valorar* por t_{N-1} ó DN (si $N \geq 30$)**

Fórmula nº 12 ANOVA-2

Factor A (muestras)

A→	1		2		3		……		k			
Factor **B** Individuos↓ o bloques	X	X^2	X	X^2	X	X^2	X	X^2	X	X^2	ΣX_A	$(\Sigma X_A)^2$
1												
2												
3												
4												
…..												
n												
											↑ **$\Sigma\Sigma X = B$** ↓	↑ **$\Sigma(\Sigma X_A)^2 = C$**
ΣX_B												
n												← **$\Sigma n = kn = N$**
$\bar{X}$												
$(\Sigma X_B)^2$												← **$\Sigma(\Sigma X_B)^2 = A$**
ΣX^2												← **$\Sigma \Sigma X^2 = D$**

$$C_A = \frac{A}{n} - \frac{B^2}{N} = \square \qquad C_T = D - \frac{B^2}{N} = \square$$

$$C_B = \frac{C}{k} - \frac{B^2}{N} = \square \qquad C_R = C_T - (C_A + C_B) = \square$$

$$V_A = \frac{C_A}{k-1} = \square \qquad V_B = \frac{C_B}{n-1} = \square \qquad V_R = \frac{C_R}{(k-1)(n-1)} = \square$$

$$Z_A = \frac{V_A}{V_R} = \square \qquad Z_B = \frac{V_B}{V_R} = \square$$

Valoración de A : por F(k-1 ; (k-1)(n-1)). **Valoración de B** : por F(n-1 ; (k-1)(n-1))

Fórmula nº 14 Valoración de r

$$Z = \frac{r\sqrt{(N-2)}}{\sqrt{1-r^2}}$$

Condiciones de aplicación para muestras pequeñas:

$V = S^2$ mayor / S^2 menor $<$ F(N-1, N-1 , 0.05)

Valoración : si N>30, por c de la D.N. y si N<30, por t_{N-2}

Fórmula nº 15: Test de rango de Spearman (r_s)

Los datos de X e Y se ordenan por separado

Individ.		X	Y	R de X	R de Y		d	d^2
1								
2								
3								
4								
5								
6								
.....								
N								
							Suma	
								Σd^2

$$r_s = 1 - \frac{6\sum d^2}{N\left(N^2 - 1\right)}$$

Z se calcula por la fórmula 14 , sin condición de aplicación, dando a r el valor de r_s

Valoración: por t_{N-2} (si N<30)
o la c de la DN (si N≥30)

*** OR (Odds Ratio)

		Enfermedad		
		+	-	
Exposición o resultado	+	a_1	a_2	Na
	-	b_1	b_2	Nb
		N_1	N_2	N

$$\mathbf{OR} = \frac{a_1 b_2}{a_2 b_1}$$

Cálculo del intervalo de confianza (IC)

Se calcula el IC del logaritmo neperiano de la OR y luego se vuelve a "números normales"... Así:

IC del ln OR= ln OR ± c $\sqrt{(1/a_1 + 1/a_2 + 1/b_1 + 1/b_2)}$

si N es menor de 30, en vez de **c** se toma **t** con g.l. N-2

luego se calculan los antilogaritmos (e^x) de los extremos del intervalo : son los límites del IC de OR. **IC =∈ (límite inferior ÷ límite superior)**

*** RR (Riesgo relativo)

$$\mathbf{RR} = \frac{a_1 N_b}{b_1 N_a}$$

Cálculo del intervalo de confianza (IC)

Se calcula primero el IC del logaritmo neperiano del RR y luego se vuelve a "números normales": IC del ln RR = ln RR ± c $\sqrt{(1/a_1 + 1/b_1 - 1/N_a - 1/N_b)}$

(¡ojo a los dos signos menos!)

si N es menor de 30, en vez de **c** se toma **t** con gl N-2

luego se calculan los antilogaritmos (e^x) de los extremos del intervalo : son los límites del IC del RR **IC =∈ (límite inferior ÷ límite superior)**

*** NNT (Número necesario para tratar)

utilizando una tabla de 2x2

	Controles	Tratados
Evento E +	a_1	a_2
Evento E -	b_1	b_2
	N_1	N_2

:

$$\mathbf{NNT} = \frac{1}{\frac{a_1}{N_1} - \frac{a_2}{N_2}}$$

*** Valoración de pruebas diagnósticas

Valoración pruebas		Resultado		
		+	-	
Enfermedad o Evento	+	**PV**	**NF**	**En**
	-	**PF**	**NV**	**San**
		P	**N**	**Total**

Sensibilidad (S) = PV*100/En
Especificidad (E) = NV*100/Sa
Eficiencia de la prueba (EP) = (PV+NV)*100/Total
Valor predictivo resultado + (VPRP) = PV*100/P
Valor predictivo resultado - (VPRN) = NV*100/N
Cociente de probabilidades + (CP+) = S/(100-E)
Cociente de probabilidades - (CP-) = (100-S)/E

P = positivo ; N = negativo
PV = positivo verdadero
NV = negativo verdadero
En = enfermos, afectados
Sa = sanos, no afectados
Si se utilizan proporciones en vez de 100 se pone 1

Algunas tasas sanitarias

Tasa de crecimiento vegetativo : 1000*(N-D)/P
Tasa de natalidad : 1000*N/P
Tasa de fecundidad general : 1000*(N/Pff)
Tasa de fecundidad por edades : 1000*(Ne/Pff)
Tasa de mortalidad general : 1000*D/P
Tasa de mortalidad infantil : 1000*($D_{<1\ año}$ / N)
Tasa de mortalidad proporcional : 100* ($D_{>50a}$ /D)

Triángulo de Pascal

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
etc.

EJERCICIOS DE BIOESTADISTICA

Dr. Eduardo Buesa
Escuela Universitaria de Enfermería "Nuestra Señora del Sagrado Corazón". Castellón

--Recordatorio de conceptos básicos (Tema 2)

1) Calcule |-2,4|*3 |3,5|*2 -4*|-8|*2 |-0,25|*(-2)

2) Calcule 5^n para valores de n = 2 , 4 , 1 , 0

3) Calcule 4! 3! 0! 1!
4) Calcule Σx para x = (8 , 5 , 3 , 2 , 6)

5) ¿Cuál es el valor absoluto de -18? ¿Y el de 5?

6) Calcule $(\Sigma x)^2$ para x = (4 , 8 , 12 , 10 , 9)

7) Calcule Σx^2 para x = (4 , 8 , 12 , 10 , 9)

8) Calcule $\binom{5}{0}$, $\binom{4}{4}$ y $\binom{4}{3}$

9) Calcule ? en e-3 * e –0.33 = e-?

10) Calcule los dos ? (un número y un signo) en e-3,25 = e-? ? e-0,25

11) Si a=3 , b=5 , c=7 , d=8 , e=2 , f=9 , g=4 , h=5 , i=5+0,00008 , j=4+k , k=∈(0÷2)
coloque estos cinco signos (no repetir!) (= , > , < , √) entre las letras:
a f d c h d f-e a+g

12) Haga lo mismo con los signos ≤ , ≥ , ≈ , ≠
g j , h i , a+b g , j g

-Medidas, redondeos, cifras significativas

50) Redondee según el consenso general los porcentajes 3,56789% , 28,1234% , 40,452%

51) Redondee según el consenso general los siguientes coeficientes de correlación:
0,12345 , 0,6785 , 0,0021 , 0,0555 , 0,99999999

52) Redondee a décimas los números: 3,335 ; 0,156 ; 6,75 ; 6,751 ; 5,55555 ; 0,05

53) Exprese con 3 decimales los números 123,456785 ; 0,0445 ; 3,141592 ; 1,000678

54) Redondee 123456 a millares

55) Calcular los límites reales de las siguientes medidas:

1,3 2 4,25 6,383 44 9,1684 2,123456

56) ¿Cuántas cifras significativas tienen estas medidas?

28 kg 0,650 g 18,300 m 0,1030 sec 0,0100 mg 165,0 meq

-Estadística descriptiva

Enunciado común para los ejercicios siguientes: En una muestra de 10 personas se determinan los componentes de la sangre X e Y, obteniendo los siguientes valores (en mg/100 ml): X: 12 , 8 , 9 , 14 , 12 , 13 , 16 , 10 , 15 , 12 Y: 9 , 15 , 10 , 8 , 7 , 11 , 13 , 12 , 10 , 9

100) Calcule la media, varianza y desviación estándar de X

101) Calcule la media, varianza y desviación estándar de Y

102) ¿Cuánto valdrían los parámetros de X si los 10 individuos constituyeran una población?

103) ¿Quién tiene más variabilidad, X o Y?

104) Calcule el coeficiente de correlación entra ambas variables

105) Calcule la ecuación de regresión de Y sobre X

106) Calcule la moda de X e Y

107) Calcule la mediana de X e Y

108) Construya una tabla de distribución de frecuencias para X

109) En una muestra la varianza vale 4 veces el cuadrado de la media. Calcule el CV

110) Se han medido dos variables cuantitativas, X y Y, en 8 individuos. Resultados:
X: 7 5 4 6 7 5 6 5
Y: 1 9 8 2 7 4 5 6
Calcular la media, varianza, CV y moda de X , la mediana de Y, el coeficiente de correlación entre X e Y y la ecuación de regresión de Y sobre X

111) La muestra A de tamaño 100 tiene un CV del 30%, el mismo que la muestra B de tamaño 50 ¿Quién tiene más variabilidad?

112) La varianza de una DB y q valen 0,8 . Escriba la notación.

--Probabilidad y sus distribuciones fundamentales

200) En el pueblo P se han construido 16 viviendas de protección oficial. Hay 32 solicitudes. Se hace un sorteo y los 4 hijos del alcalde consiguen vivienda. En el pueblo se habla de trampas. ¿Apoya la Estadística esta apreciación?

201) Se tira un dado 10 veces. Calcular la p de que la cara 3 salga 4 veces.

202) Se tiran 20 monedas. ¿Cuál es la probabilidad de que de que salgan 8 caras y 12 cruces?

203) 10 opositores han aprobado con la misma nota las pruebas para cubrir 3 plazas. Las normas prevén que en esta situación se haga un sorteo. 3 de los opositores se apellidan García. ¿Qué probabilidad hay de que saquen plaza los 3 García?

204) En una pequeña empresa los 3 empleados de la oficina quieren irse de vacaciones en agosto. El Jefe decide hacer un sorteo, poniendo en una bolsa dos bolas negras y una blanca. Quien saque la bola blanca se irá en agosto. El orden de extracción será por edad decreciente. Uno de los empleados protesta, ya que está seguro de que el primero tiene ventaja sobre los otros. ¿Está en lo cierto?

205) En la rotonda R hay una media de 4 accidentes por semana. ¿Cuál es la probabilidad de que en una semana no haya ningún accidente?

206) El Sr. X dice que adivina el resultado del lanzamiento de una moneda mientras está en el aire. En 100 lanzamientos ¿ cual es el mínimo de aciertos que apoyaría su afirmación?

207) En und DN al valor $x = 30$ le corresponde $c = 1$. La media vale 20 . Calcular s

208) El peso de las niñas de 8 años sigue una N(26'3 , 4'83) . ¿Qué peso deja por encima de él al 10% de las niñas?

209) ¿Qué puntuación original corresponde a una $c = 1,5$ en una N(10 , 2)?

210) Tipificando el valor de $x = 18$ se obtiene –0,5 . Sabiendo que la varianza vale 16, calcular la media de la distribución

211) La talla tipificada de Juanita es de –0,1 . A su edad la talla sigue una distribución normal de media 102 cm y varianza de 16 . Calcular su talla en cm.

212) La distribución del cociente intelectual (CI) de 100.000 reclutas sigue una N(95 , 10) .¿Cuántos reclutas es de esperar que tengan un CI entre 80 y 100?

213) El 75% de los chicos de 9 años de las escuelas de la provincia P tiene una talla ≥ 128 cm Sabiendo que la desviación estándar vale 6 , calcular la talla media de este colectivo

214) La probabilidad de encontrar peces de la especie X con una longitud entre 20 y 30 cm es de 0,2638 . La puntuación tipificada c de 20 es 0,4 . ¿Cuál es la notación de la distribución de la longityd de esos peces?

215) En la región R el peso medio de los recién nacidos es de 3,3 kg. con una desviación estándar de 400 g. El año pasado nacieron 5.000 niños.¿Cuantos neonatos era de esperar que pesaran 2500 g. o menos?

216) El perímetro craneal (PC) de los chicos de 13 años sigue una N(55 , 1’5) . ¿Qué porcentaje de chicos es de esperar que tengan un PC de 52 cm o menos?

217) El tiempo medio de hospitalización trás una determinada y complicada operación es de 30 días, con una varianza de 36 días. Se asume que la duración sigue la DN. En un plan de choque para acabar con las listas de espera, se asignan a la Clínica C 530 pacientes. La Administración hace cálculos para ver lo que les puede acontecer. Veamos algunos:

Número de personas en que la estancia dura 38 días o más.
La probabilidad de que un paciente esté hospitalizado al menos 25 días
La probabilidad de que un paciente esté más de 15 días ingresado, pero menos de 22
Porcentaje de personas que estén hospitalizadas como máximo 25 días.

218) Según las tablas de expectativa de vida la probabilidad de que Julio viva dentro de 20 . años es de 0’6 . Y la de su esposa María es de 0’8 . Calcular las siguientes probabilidades:

de que vivan ambos
de que sólo viva Julio
de que sólo viva María
de que no viva ninguno
de que viva uno cualquiera de los dos

219) En 80 familias de 4 hijos ¿cuántas es de esperar que tengan 2 chicos y 2 chicas? (Se asume que p=0’5)

220) En una caja tenemos 100 bolas, unas blancas y otras rojas. Sacamos repetidamente 3 bolas (reponiéndolas cada vez) y obtenemos las siguientes frecuencias de bolas blancas:

0 bolas..........343 veces ; 1 bola ....441 veces
2 bolas..........189 veces ; 3 bolas.......... 27 veces
¿Cuántas bolas blancas debe haber en la caja?

221) El 15% de los habitantes de Tordicia de Abajo va a ser agraciado en un sorteo (viaje y entrada gratis) para acompañar a su equipo de fútbol en un encuentro trascendental.¿Qué probabilidad tiene dos amigos de resultar ambos elegidos?

222) Pedro juega 20 números en un sorteo de la ONCE que da dos reintegros (para la primera y la última cifra del número). ¿Cuántos reintegros es de esperar que cobre?

223) En la fábrica F una media de 5 trabajadores acude los lunes al Consultorio médico, quejándose de molestias gástricas. ¿Cuál es la probabilidad de que un lunes no acuda nadie?

224) Una técnica operatoria tiene un 5% de complicaciones graves. ¿Cuál es la probabilidad de que en un mes en el que se realiza la técnica veinte veces haya dos complicaciones graves?

225) La duración media de la estancia hospitalaria de la enfermedad E es de 9 ± 3 días (media y desviación estándar). Calcular la p de que una estancia dure 15 días o más.

--Estimación y conformidad

250) Según las leyes de Mendel si cruzamos los genotipos AB y AB se deben obtener un 25% de AA, un 50% de AB y un 25% de BB. Se realizan 100 cruces, de los que 14 se desgracian y no pueden ser valorados. Se encuentran 21 AA , 47 AB y 18 BB. ¿Concuerda lo observado con el modelo teórico?

251) Una solución control de glucosa contiene 120 mg/dl. Probamos un glucosímetro, que según el fabricante tiene una gran precisión y exactitud, haciendo 36 determinaciones con esa solución. Se obtiene una media de 128 con una desviación estándar de 20. ¿El glucosímetro mide tan bien como se dice?

252) Estime la media poblacional de H en las chicas de A del ejercicio 324

253) ¿Cuánto vale el error estándar de G en las chicas de B del ejercicio 324?

254) Utilizando los datos del ejercicio 331, calcule el error muestral al hacer una estimación de la glucemia inicial en la población de los hombres.

255) En 15 personas con la enfermedad neurológica X se determina el cobre (Cu) en el LCR (líquido cefalorraquídeo), encontrándose una media de 16 mcg/l y una desviación estándar de de 5 mcg/l. Estimar la media de Cu en LCR en los pacientes de la enfermedad X

256) Calcular el intervalo de confianza de un coeficiente de correlación, r , que vale 0,345 obtenido de una muestra 103 individuos.

--Contraste de variables

300) Baysun y col. comparan la eficacia de midazolan (M) per os y diacepán (D) rectal para yugular ataques convulsivos en los niños (Clin. Pediatr. 2005;44:771). Para ello valoran si la convulsión cede o no antes de que pasen 10 minutos. El ataque había cedido en 18 de los 23 niños del grupo M y 17 de los 20 del grupo D. Valore los resultados.

301) Joanna Tully y col publican un trabajo en el BMJ sobre diversos factores que pueden influir en la aparición de enfermedad meningocócica en adolescentes. Recogen datos de 144 adolescentes que enfermaron y los comparan, siguiendo las normas básicas de investigación, con los de otros 144 que no enfermaron. Para la variable "besos apasionados frecuentes en la última quincena " constaba "Sí" en 42 de los que enfermaron y en 22 de los controles. Valore la relación entre besuqueos y enfermedad meningocócica.

302) En el estudio anterior otra de las variables estudiadas fue "asiste regularmente a la iglesia". Lo hacían 6 de los enfermos y 15 de los controles. Valore la relación entre esta variable y la enfermedad.

303) En la clase C hay 12 chicos y 13 chicas. Consiguen beca 4 chicos y 9 chicas. Valore estadísticamente el resultado.

304) El año 1973 hubo una epidemia de cólera en varios países del sur de Europa. En Nápoles se diagnosticaron 126 casos. De ellos 41 habían comido almejas crudas en los 5 días previos al comienzo de la enfermedad. Para aclarar si había relación entre las almejas y la enfermedad se tomaron como controles 785 personas que no enfermaron, de las que 89 habían comido almejas. ¿Conclusión?

305) En los estudios previos a la comercialización de la penicilina Chain y col publicaron en el Lancet un estudio en 48 ratones infectados con Estafilococo áureo. Se trató a 24 con penicilina, de los que murieron 3. En los no tratados no hubo supervivientes. ¿Aporta algo el experimento?

306) Pinto y col han estudiado en una zona de México la relación entre malformaciones congénitas y consanguinidad en 33194 recién nacidos en un periodo de 6 años. Hubo 1117 neonatos con alguna anomalía congénita. So tomó como control de cada caso al primer neonato del mismo sexo nacido después. 21 de los malformados tenían el antecedente de consanguinidad por 8 de los controles. Valore el resultado.

307) En una zona subdesarrollada se toman muestras de heces en busca de un determinado parásito intestinal en 200 niños de los pueblos vecinos A y B (100 de cada pueblo). En A tienen una fuente en la plaza Mayor; en B tienen que ir a buscar el agua a una fuente que dista 3 km. El parásito se encuentra en 30 niños de A y en 58 de B. ¿Es similar el grado de infestación en ambos pueblos?

308) En la región R en los últimos 20 años del siglo XIX hubo 9 veranos sin sequía. En los últimos 25 años del siglo XX los veranos sin sequía han sido 4. ¿Apoyan estos datos una modificación de las condiciones climatológicas de la región?

309) Se preguntó a 200 personas su opinión sobre el sistema sanitario público (B=bueno; R=regular; M=malo). 30 tenían un nivel de rentas alto y su opinión fue: 5 B, 10 R, 15 M. De los 70 con niveles de renta medios hubo 25 B, 30 R y 15 M. La opinión de los 100 con rentas bajas fue: 90 B, 30 R y 10 M.¿Hay diferencias en la opinión según el nivel de rentas?

310) Calcule y valore la OR correspondiente al ejercicio 300

311) Calcule y valore la OR correspondiente al ejercicio 301

312) Calcule y valore la OR correspondiente al ejercicio 302

313) Calcule y valore la OR correspondiente al ejercicio 303

314) Calcule y valore la OR correspondiente al ejercicio 304

315) Calcule y valore la OR correspondiente al ejercicio 305

316) Calcule y valore la OR correspondiente al ejercicio 306

317) Calcule y valore la OR correspondiente al ejercicio 307

318) Calcule y valore la OR correspondiente al ejercicio 308

319) Valore el coeficiente de correlación del ejercicio 104

320) ¿Hay diferencias significativas entre los valores de X e Y del ejercicio anterior?. Utilice la prueba paramétrica y la no paramétrica.

321) Cervantes y col estudian en 50 prematuros los valores séricos de LPO (que refleja el estrés oxidativo) a la edad de 4 semanas y lo relacionan con la presencia o no de retinopatía. 11 niños tenían retinopatía y la LPO dió una media de 4,36 mmol/ml, con una desviación estándar de 2,07 . Los 39 restantes no tenían retinopatía y los parámetros de su LPO fueron, respectivamente, de 2,03 y 2,40 . ¿Se explican por el azar las diferencias de LPO?

322) En el estudio del ejercicio anterior, que los autores plantean como un estudio de cohortes, se calcula la curva ROC y se encuentra que el punto de corte de 2,4 mmol/ml discrimina bien entre afectados y no afectados de retinopatía. 10 de los afectados tenían una LPO $\geq$2,4 por 12 de los no afectados. Calcule y valore el riesgo relativo de una LPO $\geq$2,4

323) Se mide el antígeno G en 10 chicas adolescentes del pueblo A y en otras 10 del pueblo B. Los resultados (en ng/ml) son los siguientes:
en A: 13 , 16 , 15 , 12 , 10 , 12 , 14 , 9 , 8 , 12
en B: 9 , 10 , 12 , 13 , 11 , 7 , 8 , 10 , 15 , 9
Valore la afirmación “Las chicas de A tienen un G claramente superior a las de B”

324) En esas mismas chicas se determina el antígeno H.
Chicas de A: 1 , 2 , 3 , 4 , 5 , 6 , 7 , 8 , 9 , 10
Chicas de B: 3 , 3 , 2 , 4 , 3 , 2 , 4 , 3 , 3 , 2
Valore la afirmación “Las chicas de A tienen un H estadísticamente similar a las de B”

325) 25 alumnos (12 chicos y 13 chicas) realizan un examen. Aprueban 4 chicos y 8 chicas. Alguien comenta: “Claro, las chicas son más aplicadas”. La Estadística ¿qué diría?

326) Se estudió el contenido de alquitrán de dos marcas de cigarrillos. En 121 cigarrillos de la marca A se encontró una media de 11 mg y una desviación estándar de 2,28 . En 181 cigarrillos de la marca B: media 13 mg, s=2 . La diferencia de 2 mg en el contenido medio de ambas marcas ¿puede interpretarse como debida al azar?

327) Los coches con catalizador ¿gastan más combustible? . Se hace una prueba con 142 coches del mismo modelo, la mitad con catalizador (C) y la otra mitad no (NC). Los C consumieron 9,8 l en media (varianza 4) y los NC 9,5 l. (varianza 4,41).

328) En la Western University de Australia se estudió la relación entre disfunción tiroidea y el colesterol de baja densidad (LD-C) expresado en mmol/l. En 580 eutiroideos la media de LD-C fue 3,5 y la desviación estándar 1 . En 30 hipotiroideos la media fue 4,1 y la desviación estándar 1,2 . La relación ¿es significativa?

329) En 10 pacientes se valora la TAS (tensión arterial sistólica) antes y después de un mes de dieta sin sal. Se quiere saber si ha sido eficaz.

Los datos:											
	antes	140	165	160	160	175	190	170	175	155	160
	después	145	150	150	160	165	175	160	165	145	170

330) Se quiere comprar un ordenador para realizar de forma reiterativa 5 programas. Se hace una prueba con tres distintos (A, que vale 700 €, B que vale 1100 y C que vale 1300) para elegir al que en conjunto sea significativamente más rápido. Si no lo hay, se elegirá el más barato. Se recoge el tiempo en segundos que cada uno tarda en ejecutar los programas :

Programa ↓	Ordenador		
	A	B	C
1	67	52	40
2	50	56	52
3	55	43	44
4	72	66	47
5	67	68	35

¿Cuál se eligió? Documente el proceso estadístico que llevó a es decisión.

331) 300 diabéticos tipo II (de ellos, 160 mujeres) se sometieron a una prueba de esfuerzo controlado de 30' de duración. Se determinó la glucemia antes y después de la prueba.

Resultados (media $\pm$ desviación estándar, en mg/dl) :

Hombres : antes 200 $\pm$ 35 ; después 148 $\pm$ 30

Mujeres : antes 190 $\pm$ 30 ; después 152 $\pm$ 40

Calculando en ambos sexos, paciente a paciente, las diferencias entre los valores antes y después se obtiene : H : 31 $\pm$ 7 ; M: 51 $\pm$ 8

Preguntas:

---1--- ¿Hay diferencias entre hombres y mujeres antes de la prueba?
---2--- ¿Hay diferencias entre hombres y mujeres después de la prueba?
---3--- ¿Hay una variación significativa en las mujeres?
---4--- ¿Hay una variación significativa en los hombres?

332) Un test de adaptación familiar realizado a 12 pares de gemelos de distinto sexo ha dado las siguientes puntuaciones:

Chicos	6	8	5	11	12	13	7	5	14	10	7	8
Chicas	7	8	10	9	11	10	16	8	15	15	9	9

Valore si hay diferencias entre chicos y chicas utilizando tanto la prueba paramétrica como la no paramétrica.

333) En el Centro de Salud de un pequeño pueblo se controlan 16 pacientes con hipertrigliceridemia. Se forman 4 grupos (A , B , C, D) y cada grupo recibe un fármaco distinto. Al cabo de dos meses se observan los siguientes descensos de triglicéridos (media y varianza): Grupo A:51 y 24,6667 ; grupo B:45 y 35,3333 ; grupo C:37 y 52,6667 ; grupo D:30,5 y 43 ¿Hay diferencias significativas entre los tratamientos?

334) Se hace un estudio con 3 detergentes (D1, D2, D3). 200 amas de casa reciben envases de D1, otras 200 de D2, otras 200 dos D3 y otras 200 de D4. Al cabo de un tiempo tienen que decir si están satisfechas con los resultados. Lo están 160 del D1, 150 del D2 y 165 del D3 . ¿Puede decirse estadísticamente que el D3 es superior a los otros?

335) Recientemente se ha publicado un estudio sobre la mortalidad de diabéticos con insuficiencia cardiaca y su relación con el control glicémico. Siguieron a un grupo de 123 pacientes, de los que murieron 36. De los fallecidos 26 tenían una HbA1c mayor del 7% por 37 de los vivos. ¿Qué se deduce de éstos datos?

336) En una región se investiga la presencia de eczema y rinitis alérgica en escolares de 11 años. Los resultados quedan reflejados en la siguiente tabla:

		Rinitis		
		si	no	
Eczema	si	141	420	561
	no	928	13.525	14.453
		1.069	13.945	15.522

Calcule y valore la asociación entre rinitis y eczema (por medio de OR).

337) En ese mismo estudio se quiere valorar la relación entre eczema y asma en esos escolares. Los resultados quedan reflejados en la siguiente tabla:

		Asma		
		si	no	
Eczema	si	249	312	561
	no	820	13.633	14.453
		1.069	13.945	15.522

Calcule y valore la OR

338) Se determinó la IgE en la sangre del cordón umbilical de 4.853 recién nacidos en el año 1995 en la provincia P, clasificándola como normal o alta. A los 5 años se vio cuántos de ellos eran asmáticos. La tabla recoge los resultados:

		Asma		
IgE		si	no	
	alta	174	50	224
	normal	12	4.617	4629
		186	4.667	4.853

Calcule y valore el riesgo relativo de ser asmático habiendo tenido la IgE alta.

339) Un equipo de profesionales del Hospital General de Boston estudiaron si la presencia de los padres influye o no en el éxito de la punción lumbar. Hubo 44 fracasos en las 281 ocasiones en que los padres no estaban presentes, mientras que fueron 1030 los éxitos en los 1178 casos con presencia de los padres. Valore los resultados.

340) El escorbuto fue durante más de dos siglos la gran plaga que azotaba a los marinos en sus viajes transoceánicos. El Dr. James Lindt hizo en 1747 un experimento en un barco con una tripulación de 12 personas, que acabó con el problema. Hizo 6 grupos de dos personas. Cada grupo recibía un tratamiento distinto. He aquí las dosis diarias de cada grupo : 1) ¼ de galón de agua 2) 25 gotas de elixir de vitriolo 3) 6 cucharadas de vinagre 4) ¼ de pinta de agua de mar 5) 2 naranjas y un limón 6) un líquido con ajo, mostaza y bálsamo del Perú. Murieron todos a excepción de los dos del grupo naranja-limón. Valore estadísticamente el contraste entre el grupo naranja-limón y el conjunto de los otros cinco tratamientos.

341) ¿Es significativo el coeficiente de correlación del ejercicio 110?

342) ¿Y el del 256?

--Valoración de pruebas diagnósticas

400) En el J Trop Pediatr de enero de 2006 se presenta un test serológico rápido para el diagnóstico de infección por Helicobacter Pylori. Se ha probado en 81 niños, a los que se hacen también las pruebas microbiológicas habituales para saber si están o no infectados. He aquí los resultados:

		Test rápido		
		+	—	
Infección	+	24	1	25
	—	3	53	56
		27	54	81

Calcular los siguientes índices de valoración de una prueba diagnóstica :

sensibilidad, especificidad y valor predictivo de resultados positivos y negativos.

¿Es útil la prueba?

401) Calcule los mismos índices para el "signo de palidez palmar" que se valoró en una región selvática de Colombia para ver si podría ser de interés en el diagnóstico de anemia. Se hizo un hemograma en 167 niños, por lo que se sabía si tenían o no anemia (48 sí , 119 no). El signo de palidez palmar fue positivo en 16 anémicos. En 95 no anémicos fue negativo. ¿Es útil?

--NNT

450) Hay indicios razonables de que los portadores del gen H sufren más ictus a partir de los 50 años que los que no lo tienen. A 300 portadores se les da todos los días una dosis del medicamento M. Otros 300 portadores sirven de control. Al cabo de 5 años se contabilizan los ictus que ha sufrido uno y otro grupo: 30 en el grupo de M y 56 en los controles. ¿A cuantas personas hay que tratar con M para evitar un ictus?

Soluciones:

1 7,2 7 -64 -0,5
2 25 625 5 1
3 24 6 1 4
4 24
5 18 5
6 1849
7 405
8 1 1 4
9 ?= 3,33
10 ?= 3 y *
11 √ > < =
12 ≤ ≈ ≠ ≥

50 3,57 28,1 40,4
51 0,123 0,678 0,056 1,000
52 3,3 0,2 6,8 6,8 5,6 0,0
53 123,457 0,044 3,142 1, 001
54 123000
55) ∈(1,25 ÷ 1,35) ; ∈(1,5 ÷ 2,5) ; ∈(4,245 ÷ 4,255) ; ∈(6,3825 ÷ 6,3835) ; ∈(43,5 ÷ 44,5) ; ∈(9,16835 ÷ 9,16845) ; ∈(2,1234555 ÷ 2,1234565)
56) 2 3 5 4 3 4

100 12,1 6,54 2,6
101 10,4 5,82 2,4
102 12,1 5,89 2,4
103 Y, pues su CV es mayor : 23,1% contra 21,5%
104 -0,277
105 y = 13,564-0,261x
106 X: 12 Y: 9 y 10
107 12 10
108 Siguiendo los pasos previstos:

1) Valor más alto 16, el más bajo 8 . R = 16,5 – 7,5 = 9
2) Nº clases = 4
3) I = 9/4 = 2,25 → 3
4) Sobras: (4*3)-9 = 3 . Reparto: 2 al inicio, 1 al final (podría ser al revés)
5) La 1ª clase comenzará en 7,5–2=5,5 (LR) ó en 6 (LT) la última terminará en 16,5+1=17,5 (LR) ó en 17 (LT)
6) Construimos la tabla completa

LT	LR	p.m.	f	f.r.	%	∑f	∑fr	∑%
6 - 8	5,5 - 8,5	7	1	0,1	10	1	0,1	10
9 - 11	8,5 - 11,5	10	2	0,2	20	3	0,3	30
12 - 14	11,5 - 14,5	13	5	0,5	50	8	0,8	80
15 - 17	14,5 - 17,5	16	2	0,2	20	10	1	100
			10	1	100			

7) La tabla para publicar podría ser

Valores de X		
mg/100 ml	n	%
6 - 8	1	10
9 - 11	2	20
12 - 14	5	50
15 – 17	2	20

109 $\mathbf{s^2 = 4\bar{x}^2 \;\; ; s = 2\bar{x} \; ; CV = \frac{2\bar{x}}{\bar{x}}\% = \; 2\%}$

110 5,6 ; 1,12 ; 18,9% ; 5 ; 5,5 ; r = -0,538 ; y = -1,4x +13,3

111 tienen ambos la misma variabilidad, al ser el mismo el CV

112 s2 = npq; p= 0,2 ; 0,8= n*0,2*0,8 ; n= 5 Es B(5 , 0,2)

200 Es B(4 , 0'5) ; p(r=4) = 0,0625 ó 6,25% . La probabilidad es baja, pero posible. La estadística no apoya la apreciación de fraude, pero tampoco la excluye...
Por la ley multiplicativa se obtiene el mismo resultado: 0,5*0,5*0,5*0,5=0,0625

201 B(10 , 1/6) ; p(r=4) = 0,0543

202 B(20 , 0,5) ; p(r=8) = 0,1201 Lo mismo da p(r=12) = 0,1201

203 B(3 , 0'3) ; p(r=3) = 0,027 ó 2,7%

204 Los 3 tienen la misma probabilidad
---saca el 1º : p(A) = 1/3 ; si acierta, gana
---no acierta y saca el 2º, cuya p es : p(B/A)= 2/3 * 1/2 = 1/3 ; si acierta, gana
---no acierta y saca el 3º (que acertará por fuerza) : p(C/AyB)= 2/3 * 1/2 *1 = 1/3

205 La variable accidente es de entrada binomial. Sólo nos da la media (que es menor de 5), por tanto tiene que ser una D de Poisson. P(0,4) ; p(r=0) = 0,6703 ó 6,70%

206 B(100, 0'5) $\rightarrow$ N(50 , 5)
para $\alpha = 0{,}05$ el límite de significación está en 1,96 . Hay que hallar el valor original correspondiente a c=1,96 , que nos dará el mínimo de aciertos para que sus pronósticos sean significativos. 1,96 = (x-50)/5 ; x = 59,8 ≈ 60 , que es el mínimo de aciertos que apoyarían su afirmación a un nivel de significación de 0,05 .
(Si se toman niveles más estrictos el nº aumenta: 63 para $\alpha = 0{,}01$ y 66 para $\alpha = 0{,}001$)

207 Aplicando la fórmula de c: s=10

208 Problema de campana. Hay que calcular primero la c que deja por debajo de ella el 40% de superficie (50-10), o sea una p de 0,4 (0,5-0,1). Le corresponde una c=1,28 . Aplicando la fórmula de c, x=32,5 kg

209 Aplicando la fórmula de c , x=13

210 De forma similar , $\bar{\mathbf{X}}$=20

211 " " , X=101,6

212 Problema de campana: c_1= -1,5 ; c_2= 0,5 ; p(80≤x≤100) = 0,6247

213 para p de 0,25 c=-0,67 ; por la fórmula de c , $\bar{\mathbf{X}}$=132 cm

214 N(16 , 10) : a una c de 0,4 corresponde una p de 0,1554 ; 0,1554+0,2638=0,4192 , al que corresponde una c=1,4 ; 0,4=(20- $\bar{\mathbf{X}}$)/s ; 1,4=(30- $\bar{\mathbf{X}}$)/s ; resolviendo estas 2 ecuaciones con dos incógnitas: $\bar{\mathbf{X}}$=16 y s=10

215 Es una N(3'3 , 0'4). Se pide Ni para x≤2'5 . Se calcula p y se mutiplica por 5000; Ni = 114

216 0,0228 ó 2,28%

217 49 0,7967 0,0856 20,3%

218 0,48 ; 0,12 ; 0,32 ; 0,08 ; 0,92

219 Es B(4 , 0’5). Nr = 80 * p(r=2) = 30
220 Se calcula la notación, que es B(3 , 0’3) ; 100 *0,3 = 30
221 0,0225
222 Es la media o esperanza matemática de una B(20 , 0’2) = 20*0,2=4
223 Es P(5) ; p(r=0) = 0,06738
224 Es B(20 , 0’05) ; p(r=2) : 0,1839
225 Es N(9 , 3) ; se pide p(x≥15) ; p= 0,0228

250 Problema de conformidad entre frecuencias observadas y las teóricas. H_0 : no hay diferencias entre ambas, hay conformidad. → fórmula nº 3

	AA	AB	BB	
observado	21	47	18	86
esperado	21,5	43	21,5	86

Z=0,953 < (2 , 0’05)=5,99
no se puede rechazar H_0 , por tanto hay concordancia

251 Problema de conformidad de una media observada con otra teórica. H_0 : es conforme, no hay diferencias entre lo observado y lo teórico

$\mathbf{Z = \frac{(128-120)\sqrt{36}}{20} = 2,4}$; Z > $c_{0,05}$ = 1,96 con rechazo de H_0 a ese nivel.

No hay conformidad, el glucosímetro no mide tan bien...

252 al ser muestra pequeña no es posible una estimación puntual
estimación por intervalo: $\mathbf{IC\overline{x} = 5,5 \pm 2,16 \text{ ó } \in (3,34 \div 7,66)}$

253 $\mathbf{e = \frac{0,7379}{\sqrt{10}} = 0,233}$

254 $\mathbf{E = c * e = 1,96 \frac{35}{\sqrt{140}} = 5,80}$

255 IC de la media: 16 ±8 ó ∈(13,2 ÷ 18,8)
256 IC de r = ∈(0,162 ÷ 0,505)

300 Contraste frecuencias 2 Vbles. CL con 2 modalidades cada una: medicamento (M y D) y cese ataque (Si, No). Datos independientes. A resolver por fórmula 2 (ó 1) . Ho: las frecuencias de éxito de ambos medicamentos son similares; no hay dif. significativas. Tabla 2x2:

Cese ataque

		Si	No	
Medic	M	18	5	23
	D	17	3	20
		35	8	43

La E de b2 vale 3,72 .Como está entre 3 y 5 → Yates
Zc = 0,0301 < χ^2 (1, 0’05) = 3,84 → no rechazo H_0 , n.s.
La eficacia de ambos medicamentos es similar

301 Fórm. 2
cumple condición aplicación
Z=8,04 > $\chi 2$ (1, 0'01)=6,64
→ rechazo H_0 a ese nivel de significación. Sentido: hay relación significativa entre besuqueos y enfermedad, es un factor de riesgo

	Mening.			
		si	no	
Be	si	42	22	64
sos	no	102	122	224
		144	144	288

302 Fórm. 2
cumple condición aplicación
Z=4,16 > $\chi 2$ (1, 0'05)=3,84
→ rechazo H_0 a ese nivel de significación. Sentido: los que van a la iglesia enfermaron menos, es un factor protector.

	Mening.			
		si	no	
Igle	si	6	15	21
sia	no	138	129	267
		144	144	288

303 Fórm. 2
cumple condición aplicación
Z=3,22 < $\chi 2$ (1, 0'05)=3,84
→ no rechazo H_0 . La diferencia observada aún se puede explicar por el azar.

	Sexo			
		chicos	chicas	
Apro	si	4	9	13
bar	no	8	4	12
		12	13	25

304 Fórm. 2
cumple condición aplicación
Z=39,9 > $\chi 2$ (1, 0'001)=10,83
→ rechazo H_0 a ese nivel de significación. Sentido: hay relación significativa entre la ingestión de almejas y la enfermedad

	Cólera			
		si	no	
alme	si	41	89	130
jas	no	85	696	781
		126	785	911

305 Fórm. 2
cumple condición aplicación
Z=37,33 > $\chi 2$ (1, 0'001)=10,83
→ rechazo H_0 a ese nivel de significación. Sentido: hay relación significativa entre la penicilina y el sobrevivir. Con penicilina menos muertos

	Penicilina			
		Si	No	
Muerte	si	3	24	27
	no	21	0	21
		24	24	48

306 Fórmula 2
cumple condición aplicación
Z=5,904 > $\chi 2$ (1, 0'02)=5,41
→ rechazo H_0 a ese nivel de significación. Sentido: hay relación significativa entre consanguinidad y malformac.
Hay más malformaciones entre consanguíneos.

	Malformac			
		Si	No	
Consang.	si	21	8	29
	no	1096	1109	2205
		1117	1117	2234

307 Fórm. 2
cumple condición aplicación
Z=15,9 > $\chi 2$ (1, 0'001)=10,83
→ rechazo H_0 a ese nivel de significación. Sentido: hay más parásitos en B. El grado de ingestación no es similar

		Pueblo		
		A	B	
Pará sitos	si	30	58	88
	no	70	42	112
		100	100	200

308 Fórm.2
cumple condición aplicación
Z=4,55 > $\chi 2$ (1, 0'05)=3,84
→ rechazo H_0 a ese nivel de significación. Sentido: en el siglo XX menos veranos sin sequía (o más con sequía). Sí apoyan estos datos el cambio climático, pero no lo "demuestran", pues es estudio de observación y no experimental.

		Siglo		
		XIX	XX	
Sequía	si	11	21	32
	no	9	4	13
		20	25	45

309 Fórm. 3
cumple condición aplicación
Z=32,24 > $\chi 2$ (4, 0'001)=18,47
→ rechazo H_0 a ese nivel de significación.
Sentido: relación inversa entre renta y satisfacción
Los de renta alta valoran mejor y los de renta baja peor

		Renta			
		A	M	B	
Opinión	B	5 (13,5)	25 (31,5)	60 (45)	90
	R	10 (10,5)	30 (24,5)	30 (35)	70
	M	15 (6)	15 (14)	10 (20)	40
		30	70	100	200

310 OR = 0,635 ; IC = $\in$(0,131÷3,076) ; n.s. , pues incluye el 1
311 OR = 2,28 ; IC = $\in$(1,28÷4,07) ; no incluye el 1 → rechazo H_0 , p<0,05 (es el nivel de significación empleado) .Sentido: hay relación + entre besuqueos y meningitis
312 OR = 0,374 ; IC = $\in$(0,141÷0,993) ; no incluye el 1 → rechazo H_0 , p<0,05 (es el nivel de significación empleado) .Sentido: hay relación negativa entre visita iglesia y meningitis (los que van, menos meningitis)
313 OR = 0,222 ; IC = $\in$(0,038÷1,311) ; n.s. , pues incluye el 1
314 OR = 3,772 ; IC = $\in$(2,446÷5,817) ; no incluye el 1 → rechazo H_0 , p<0,05 (es el nivel de significación empleado) .Sentido: hay relación + entre consumo de almejas e infección
315 Como hay una casilla con valor 0, no se puede calcular la OR. Hay una solución aproximada, que es sumar 0,5 a cada casilla. Nos quedan: 3,5 , 24,5 , 21,5 , 0,5
OR = 0,003 ; IC = $\in$(0,000÷0,068) ; no incluye el 1 → rechazo H_0 , p<0,05 (es el nivel de significación empleado) .Sentido: hay relación negativa entre penicilina y muerte, menos muertos en los que recibieron penicilina
316 OR = 2,656 ; IC = $\in$(1,171÷6,022) ; no incluye el 1 → rechazo H_0 , p<0,05 (es el nivel de significación empleado) .Sentido: hay relación + entre consanguinidad y malformaciones
317 OR = 0,310 ; IC = $\in$(0,173÷0,556) ; no incluye el 1 → rechazo H_0 , p<0,05 (es el nivel de significación empleado) .Sentido: en A hay menos ingestaciones
318 OR = 0,233 ; IC = $\in$(0,058÷0,930) ; no incluye el 1 → rechazo H_0 , p<0,05 (es el nivel de significación empleado) .Sentido: en el siglo XIX menos veranos de sequía. Sí apoya
319 r = -0,277 , N = 10 , cumple condición aplicación: V=1,124 < F(9 , 9 , 0'05)=3,18 , Z = -0,815 , |Z| < t(8 , 0'05)=2,306 n.s. No hay relación significativa

320 El contraste de dos Vbles CT puede ser enfocado como tal , pero también como el contraste de una variable CL, análisis, con dos modalidades, X e Y, y otra CT, el resultado del análisis , con datos apareados.

Fórmula nº 10 :

$\overline{x}_d = 1,7 \quad s_d^2 = 15,79$

$Z = 1,35 < t(9 , 0,05) = 2,262$ n.s.

Test de los signos:

$N_1 = 7$, $N_2 = 3$, $N = 10$

$x = 7$, $Z = 1,265 < t(9 , 0,05) = 2,262$

n.s.

321 Fórm. 6, Cumple: V=5,76/4,2849=1,344 <F(10 , 38 , 0'05) ≈ 2
s^2 común = 5,4527 ; Z=2,923 > t(48 , 0'01) ≈ 2,689 ; p<0,01 Sentido: LPO más alta si hay retinopatía. Las diferencias observadas no se explican por el azar.

322 Tabla de 2x2 :

RR = **12,7**272727 ; su ln=2,54374715

IC ln = ∈(0,565378÷4,52215624)

y por tanto IC de RR = ∈(1,76÷92,03) ;

no incluye el 1 → rechazo H_0 ; la LPO alta se asocia significativamente con retinopatía, el punto de corte discrimina bien.

	Retinopatía			
		Si	No	
LPO	≥2,4	10	12	22
	<2,4	1	27	28
		11	39	50

323 Fórm. 6 , cumple condición, Z=1,529 < t(18 , 0'05)=2,101, n.s.
No puede mantenerse la afirmación

324 Fórm. 6 , no cumple condición aplicación (V=16,83 > F(9 , 9 , 0'05)=3,18) → Mann-Whitney
R1 = 131 , R2 = 79 , Z = -1,9654 < t(18 , 0'05)=2,101 n.s. No puede rechazarse la afirmación

325 Fórm. 2 , cumple , Z = 1,989 < $\chi 2$ (1 , 0'05) =3,84 n.s. No hay diferencias significativas entre chicos y chicas

326 Fórm 6 , s^2 común = 4,479 , Z = -8,05 > $c_{0,001}$=3,30 → rechazo H_0 : p<0,001
En B es significativamente más alto, la diferencia no se debe al azr

327 Fórm. 6 , s^2 común = 4,205 , Z = 0,872 n.s. "Gastan igual"

328 Fórm. 6 , s^2 común = 1,021 , Z = -3,17 > $c_{0,01}$ = 2,58 → rechazo H_0 . Sí hay relación significativa, los hipotiroideos tienen el LD-C más alto

329 Vbles: CL :momento (antes , después) y CT: TAS , datos apareados , Fórm. 10

$\overline{x}_d = 6,5 \quad s_d^2 = 72,5$

$Z = 2,414 > t(9 , 0'05) = 2,262$

p<0,05 , sentido: más baja después

Ha sido eficaz

330 ANOVA-1 , Z = 5,90 > F(2 , 12 , 0,05)= 3,89 → rechazo H_0 → Scheffé
A y C: Z_{sch} = 5,72 p<0'05 , sentido: C es más rápido
A y C: Z_{sch}=0,43 n.s. B y C : Z_{sch}=3 n.s.
Como se elige al más rápido, le elección recae en C, aunque sea el más caro.

331 -1- fórm. 6 , Z=2,665 , $Z>c_{0,01}=2,58 \rightarrow$ rechazo H_0: Hay diferencias, en hombres más alto
-2-Fórm. 6 , Z= -0,9686 $>c_{0,05}=1,96$ n.s.
-3-datos apareados, Fórm. 10 , Z=80,64 p<0,001 , Sí la hay
-4-datos apareados, Fórm. 10 , Z=52,39 p<0,001 , Sí la hay

332 prueba paramétrica: Fórm. 6 , cumple, Z = -1,392 n.s. No hay diferencias
prueba no paramétrica, Mann-Whitney: R1 = 125,5 , R2 = 174,5 Z= -1,495 n.s. No difer.

333 ANOVA-1 , No se dan los datos originales, sino parámetros. se empieza a mitad de la plantilla...
Z= 8,30 > F(3, 12, 0'01)=5,95 rechazo $H_0 \rightarrow$ Scheffé
Sólo son significativos: A y D (Zsch = 7,20 , p<0,01) y B y D (Zsch=3,60 , p<0,05).
Sí hay diferencias significativas entre los tratamientos, a expensas de valores más altos en A y B y más bajos en D

334 Fórm. 3 , cumple, Z=3,54 < $\chi 2$(2 , 0'05)=5,99 , n.s. No puede decirse

335 Fórm. 2 , cumple. Z =8,99 p<0,01
La HbA1 baja va asociada a una menor mortalidad

	Muerte			
		Si	No	
HbA1c	>7%	26	37	63
	≤7%	10	50	60
		36	87	123

336 OR = 4,89 IC = $\in$(3,998÷5,988) , no incluye el 1 $\rightarrow$ rechazo H_0 , p<0,05 .Rinitis y Eczema se asocian significativamente.

337 OR = 13,3 IC = $\in$(11,073÷13,899) ; no incluye el 1 $\rightarrow$ rechazo H_0 , p<0,05 . Hay asociación significativa entre eczema y asma

338 RR = 299,65 IC = $\in$(169,6÷529,5) ; no incluye el 1 $\rightarrow$ rechazo H_0 , p<0,05 La IgE alta conlleva un riesgo alto de ser asmático

339 Se puede resolver por la Fórm. 2 o por OR
---Form.2- Z=1,901 n.s. La presencia de los padres no influye en el éxito o fracaso de la punción
--- OR = 0,774 IC = $\in$(0,537÷1,115) , que incluye el 1 y por tanto n.s. En presencia de los padres hay menos fracasos, pero esto no llega a ser significativo.

		P.L.		
		Fracaso	Exito	
Presencia padres	Si	148	1030	1178
	No	44	237	281
		192	1267	1459

340 Fórm. 2 , pero no cumple, pues $a_1 = 1{,}67 \rightarrow$ p de Fisher .

	Tratamiento			
		Naranja limón	Otros	
Muerte	Si	0	10	10
	No	2	0	2
		2	10	12

$$\frac{N1!N2!Na!Nb!}{N} = 109963{,}6364$$

a1!a2!b1!b2! = 7257600

pF = 0,015 (unilateral)

0,030 (bilateral)

A pesar de la escasa muestra la diferencia es significativa a favor del grupo naranja-limón

$p < 0{,}05$

341 No. No se puede aplicar la Fórmula nº 14, pues V supera a la F de referencia. Hay que calcular y valorar la r de Spearman, que vale –0,476 . Z = -1,326 < t(6, 0'05)= 2,447 ; n.s.

342 Fórmula 14 ; Z= 3,69 ; p<0,001

400 S=96% , E=94,6% , VPRP = 98,1% , VPRN=98,1% La prueba es útil ya que S y E son altas, mayores del 90%

401

S=33,3% , E= 79,8% , VPRP=40% , VPRN= 74,8%

No es una buena prueba, pues S es muy baja. Tiene cierto valor si sale negativa.

	Signo			
		+	-	
Anemia	Si	16	32	48
	No	24	95	119
		40	127	167

450 NNT = 11,538... ≈ 11 ó **12**